UNIVERSIDADE FEDERAL DO PARANÁ

NICOLE AMANDA ROZIN

PREVISÃO DO DESLOCAMENTO DE TEMPESTADES SEVERAS: ABORDAGENS POR APRENDIZADO DE MÁQUINA

CURITIBA

2018

NICOLE AMANDA ROZIN

PREVISÃO DO DESLOCAMENTO DE TEMPESTADES SEVERAS: ABORDAGENS POR APRENDIZADO DE MÁQUINA

Dissertação apresentada ao Curso de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação Matemática, do Departamento de Matemática, Setor de Ciências Exatas e do Departamento de Construção Cívil, Setor de Tecnologia, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Métodos Numéricos em Engenharia.

Orientador: Prof. Dr. Paulo Henrique Siqueira

Coorientador: Dr. Cesar Augustus Assis Beneti

CURITIBA

2018

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

R893p Rozin, Nicole Amanda
Previsão do deslocamento de tempestades severas: abordagens por aprendizado de máquina / Nicole Amanda Rozin. – Curitiba, 2018.

Dissertação - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Métodos Numéricos em Engenharia, 2018.

Orientador: Paulo Henrique SiqueiraCesar Augustus Assis Beneti.

1. Tempestades. 2. Aprendizado de máquinas. 3. Análise de regressão. 4. Modelos lineares (Estatística) . I. Universidade Federal do Paraná. II. Siqueira, Paulo Henrique. III. Beneti, Cesar Augustus Assis. IV. Título.

CDD: 551.55

Bibliotecário: Elias Barbosa da Silva CRB-9/1894

MINISTÉRIO DA EDUCAÇÃO
SETOR CIÊNCIAS EXATAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO MÉTODOS NUMÉRICOS
EM ENGENHARIA

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **NICOLE AMANDA ROZIN** intitulada: **PREVISÃO DO DESLOCAMENTO DE TEMPESTADES SEVERAS: ABORDAGENS POR APRENDIZADO DE MÁQUINA**, após terem inquirido a aluna e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 03 de Agosto de 2018.

PAULO HENRIQUE SIQUEIRA
Presidente da Banca Examinadora (UFPR)

LEONARDO CALVETTI
Avaliador Externo (UFPEL)

LUZIA VIDAL DE SOUZA
Avaliador Interno (UFPR)

Dedicado aos meus pais,
Andreia e Ricardo.

AGRADECIMENTOS

A Deus, por tudo.

Aos meus pais, Andreia e Ricardo, por serem a minha inspiração, por me ensinarem a persistir diante das dificuldades, acreditarem em mim e estarem sempre ao meu lado.

Ao meu noivo, Junior Hervis, pela sua paciência, apoio, motivação, carinho e dedicação diante dos momentos de maiores tensões.

À minha irmã, Nikele, e meu cunhado, Alex, pelos ouvidos atenciosos nos momentos de maiores anseios.

À minha família, pelo carinho e apoio, especialmente, aos meus tios Ivan e Leila, meus primos Leomar, Sandra e Jessica, por me proporcionarem mais que uma residência: um lar.

Aos meus orientadores, Paulo Siqueira e Beneti, pela orientação, ensinamentos, contribuições e paciência, sem os quais não seria possível a conclusão deste trabalho.

Aos meus amigos, pela compreensão e apoio. À todos os colegas e amigos do PPGMNE e do SIMEPAR, especialmente, ao Bruno Rutyna, Camila Oliveira, Diandra Kubo, Inajara Rutyna, Ivan Guimarães, Jorge Bonato, Lucas Lamy, Mauren Micalichen e Michely Oliveira que foram fundamentais nessa jornada.

Ao meu amigo, Guilherme Pianezzer, por sua amizade e confiança que possibilitaram essa oportunidade.

Ao SIMEPAR e ao PPGMNE pela oportunidade e por todo o aprendizado proporcionado por meio de profissionais experientes e dedicados.

Ao professor Leonardo Calvetti e à professora Luzia Vidal, pela honra de aceitarem compor a banca examinadora e por suas valiosas contribuições.

A todos, que de alguma forma contribuíram para a realização deste trabalho, meus sinceros agradecimentos.

RESUMO

A previsão de tempestades severas pode auxiliar no processo de tomada de decisão e nas medidas operacionais, bem como ajudar a mitigar e até mesmo antecipar os danos, permitindo que as ações possíveis sejam tomadas. Portanto, existe a necessidade de técnicas confiáveis e rápidas para o monitoramento de tempestades, que consiste em três processos principais: identificação de células de tempestades ativas, rastreamento, e também a previsão de seu deslocamento. O foco deste trabalho é o terceiro passo, com o objetivo de estudar métodos de aprendizado de máquina para previsão de tempestades de curto prazo em células identificadas e rastreadas pelo sistema TITAN (Identificação, Rastreamento, Análise e Previsão de Tempestades) em diferentes estágios. A análise ocorre na região sul e sudeste do Brasil e usa dados de radares meteorológicos e descargas elétricas atmosféricas. Devido à natureza dos fenômenos representados neste trabalho, métodos de aprendizado de máquina foram escolhidos porque eles são capazes de entender e aprender com os recursos e seus relacionamentos. Além disso, uma vez que o modelo é aprendido pelo método escolhido, o processamento das novas entradas ocorre rapidamente. Dois tipos de técnicas de regressão são estudadas: Ensemble e Modelo Linear. Foram aplicados os seguintes métodos para a previsão: Bagging, Random Forest, Extra Trees, Theil Sen e Bayesian Ridge. A avaliação dos resultados é feita comparando-os com a previsão fornecida pelo TITAN para cada célula, uma vez que é uma ferramenta bem estabelecida na área. O melhor desempenho foi obtido com o Algoritmo Random Forest. Seus resultados mostraram-se satisfatórios para a predição de deslocamento, mostrando-se uma boa alternativa ao software padrão. Além disso, uma contribuição mais evidente dos métodos propostos é encontrada para a previsão do tamanho das tempestades.

Palavras chaves: Aprendizado de Máquina. Regressão. Previsão de Tempestades. Aprendizado Agrupado. Modelo Linear.

ABSTRACT

Thunderstorm forecast can help in the decision-making process and operational measures, as well as help mitigate and even anticipate damage, allowing those decision to be taken. Therefore, there is a need for trustworthy and fast techniques for storms monitoring, consisting of three main processes: identification of active storm cells, tracking, and also their forecast their displacement. The focus of this work is the third step, aiming to study machine learning methods for short-term storm forecast on cells identified and tracked by TITAN (Thunderstorm Identification, Tracking, Analysis, and Nowcasting) system in different stages. The analysis takes place in the discussed region and uses data from meteorological radars and atmospheric electrical discharges. Due to the nature of the phenomena represented in this work, machine learning methods are chosen because they are able to better understand and learn from the features and their relationships. Moreover, once the model is learned by the chosen method, the processing of the new entries occurs fastly. Two types of regression techniques are studied: Ensemble and Linear Model. In totally, it was applied the following methods for the forecast: Bagging, Random Forest, Extra Trees, Theil Sen and Bayesian Ridge. The evaluation of the results is done by comparing them with the forecast provided by TITAN for each cell, since that is a well-established tool in the area. The best performance was achieved with the Random Forest Algorithm, and its results proved to be satisfactory for the prediction of displacement, shown to good alternative to the standard software. In addition, a more evident contribution of the proposed methods was found to the prediction of the storms' shape.

Keywords: Machine Learning. Regression. Thunderstorm Forecasting. Ensembles. Linear Models.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E DE SIGLAS

AM – Aprendizado de Máquina
BRE – *Bayesian Ridge Estimator*
CART – *Classification and Regression Trees*
CEMIG – Companhia Energética de Minas Gerais
EMA – Erro Médio Absoluto
ETE – *Extra Trees Estimator*
INPE – Instituto Nacional de Pesquisas Atmosféricas
MMQ – Método dos mínimos quadrados
REMQ – Raíz do Erro Médio Quadrático
RFE – *Random Forest Estimator*
RINDAT – Rede Integrada Nacional de Detecção de Descargas Atmosféricas
SEQ – Soma do Erro Quadrático
SIMEPAR – Sistema Meteorológico do Paraná
TBE – *Tree-Bagging Estimator*
TITAN – *Thunderstorm Identification, Tracking, Analysis and Nowcasting*
TSE – *Theil Sen Estimator*
UTC – *Universal Time Coordinated*
VIL – *Vertically Integrated Liquid-Water*

LISTA DE SÍMBOLOS

d – Distância do centróide previsto ao observado
D – Conjunto de dados
D_t – Conjunto de teste
D_T – Conjunto de treinamento
k – Número de características de uma amostra
n – Número de exemplos no conjunto de dados
w – Vetor de coeficientes da regressão linear
x – Vetor de atributos
X – Matriz de atributos
y – Valor esperado
$\hat{y}$ – Valor estimado

Símbolos Gregos Minúsculos

ϵ – Resíduo de regressão
μ – Média das distâncias
σ – Desvio padrão das distâncias

SUMÁRIO

1 INTRODUÇÃO

O sul e sudeste do Brasil são regiões propícias à ocorrência de tempestades severas. Visto que a principal atividade econômica do país é a agroindústria, setor vulnerável a precipitação e eventos relacionados, esse tipo de evento pode afetar a economia além de apresentar riscos à vida (BENETI, 2012).

Essa mesma região é responsável por cerca de um terço da produção de energia elétrica do país. De acordo com o INPE (2015 apud KLEINA, 2015), grande parte dos desligamentos e problemas de distribuição no Brasil, respectivamente 70% e 40%, no setor de energia elétrica são ocasionados por descargas atmosféricas.

Devido ao clima tropical e subtropical, o Brasil é líder mundial em ocorrência de descargas atmosféricas. Além de apresentarem riscos ao setor elétrico e outros setores econômicos, ainda representam riscos à vida. Segundo estatística apresentada pelo INPE (2017), foram registradas 1.792 mortes no Brasil, entre os anos de 2000 e 2014, e a taxa de ocorrência de raios é de 50 milhões por ano no país.

O monitoramento de tempestades severas, principalmente quando existem riscos envolvidos, pode auxiliar na tomada de decisões e de medidas operacionais, uma vez que permite conhecer o deslocamento e o comportamento de tais tempestades.

Quando se monitora as tempestades elétricas, diversos atributos “podem ser estimados e utilizados para categorizar um evento atuante” (KLEINA, 2015). Assim, surge a necessidade de buscar técnicas de previsão rápidas e confiáveis, a fim de melhorar a qualidade de alertas de eventos meteorológicos (BONATO, 2014).

As informações obtidas de radar meteorológico, de acordo com Damian (2012), possibilitam “estimar com mais precisão quais foram os eventos que deram origem ao fenômeno meteorológico e qual será seu comportamento no futuro”.

Han et al. (2009), reconhece a contribuição do uso de dados de radares meteorológicos no processo de acompanhamento de tempestades e ressalta que são "importantes na previsão da localização e da força dos eventos climáticos severos".

Nesse contexto, o monitoramento referido trata de três etapas, a identificação de células ativas de tempestades, o acompanhamento de seu deslocamento ao longo do tempo, bem como sua análise e previsão a curtíssimo prazo.

Previsões imediatas de eventos dessa natureza, segundo Wilson et al. (1998), "são particularmente importantes para a aviação comercial e geral, eventos esportivos ao ar livre, indústria da construção, utilitários elétricos e transporte terrestre". Permitem mitigar e até mesmo antecipar danos, permitindo que as ações possíveis possam ser

tomadas.

Neste trabalho, tem-se por preocupação a terceira etapa e objetiva-se avaliar diferentes tipos de modelos de Aprendizado de Máquina para a previsão do deslocamento de tempestades severas comparados ao software TITAN (*Thunderstorm Identification, Tracking, Analysis and Nowcasting*) (DIXON; WIENER, 1993).

Alguns objetivos específicos permeiam o desenvolvimento desse trabalho. Tratam da necessidade de conhecer os dados, assim como as técnicas e estudá-las para obter um modelo bom e aplicável ao problema. Esses objetivos estão listados na subseção 1.1.2.

Assim, tem-se por foco o estudo de técnicas de Aprendizado de Máquina (AM) aplicadas a previsão desses fenômenos, a curtíssimo prazo.

Para isso se propõe o uso dos dados de radar meteorológico e de descargas atmosféricas, visto que contemplam diversas informações dos eventos de interesse e, por isso, se fazem ferramentas relevantes aos objetivos que a pesquisa se propõe.

Para esse fim, as primeiras etapas são obtidas através da ferramenta TITAN, na qual são extraídos os dados de células de tempestades identificadas e rastreadas em diferentes estágios de vida.

A proposta se estende para a região sul e sudeste do Brasil e se restringe a tempestades que tenham sido processadas, identificadas e acompanhadas pelo software no período de agosto de 2016 a janeiro de 2018. Ou seja, não se limita apenas ao desempenho do software nesses processos mas, também, ao controle de qualidade realizado em cada um dos radares que compõem o mosaico de dados utilizado por ele.

Devido à natureza dos fenômenos, os métodos de Aprendizado de Máquina são propostos, pois apresentam um grande potencial em entender e aprender melhor com as características e relacionamentos dos dados. Além disso, uma vez que o modelo é aprendido pelas ferramentas de AM, com a confiabilidade desejada, o processamento das novas entradas ocorre rapidamente.

A avaliação dos modelos propostos é feita por diferentes métricas de mensuração de erro e discutidos em comparação com a previsão fornecida pelo TITAN para cada célula, pois trata-se de uma ferramenta reconhecida e estabilizada na área.

Com esse estudo é proposto um modelo alternativo ao software, além de estender a avaliação dos algoritmos a outras atributos relativos ao tamanho da tempestade.

Diversas técnicas para regressão são testadas, todas utilizando características das tempestades, como a posição, orientação, histórico de deslocamento, tamanho dos eixos, VIL, número de raios, entre outras.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Estudar técnicas de Aprendizado de Máquina para a previsão do deslocamento de tempestades severas a curtíssimo prazo.

1.1.2 Objetivos específicos

- Selecionar dados de radares meteorológicos e descargas elétricas para comporem atributos dos modelos de Aprendizado de Máquina;
- Analisar o desempenho de diferentes métodos de Aprendizado de Máquina aplicados a previsão do deslocamento de tempestades;
- Avaliar o a extensão dos modelos propostos para a previsão de outros atributos da tempestade;
- Verificar a contribuição dos modelos de Aprendizado de Máquina em melhoria a previsão proveniente do software TITAN;
- Propor um modelo de previsão de deslocamento de tempestades alternativo ao software TITAN.

1.2 ESTRUTURA DO TRABALHO

Em um primeiro momento, no Capítulo 2, são apresentados os conceitos meteorológicos, que consistem da definição e apresentação da previsão de tempestades a curtíssimo prazo, bem como o acompanhamento, os radares meteorológicos, seus dados e as descargas elétricas atmosféricas. Na sequência, define-se o TITAN e cada um dos seus processos.

Os conhecimentos relativos ao AM, são dissertados no Capítulo 3. Desde as principais definições e tipos de problemas, assim como os algoritmos da pesquisa que se dão em dois grupos: *ensemble* e modelo linear.

Do primeiro grupo, são expostos, na seção 3.3, os métodos *Bagging*, *Random Forest* e *Extra Trees*, todos baseados no aprendizado agrupado por aleatorização no conjunto de árvores de decisões.

Já para o segundo, discorre-se na seção 3.4 quanto aos algoritmos *Theil Sen* e *Bayesian Ridge*, que tratam-se de modelos lineares robustos, ou seja, resistentes a dados ruidosos.

Delineados os dados e métodos da pesquisa, no Capítulo 4 são explanadas toda a manipulação necessária para a construção dos modelos propostos, desde a

configuração e seleção de características, a união dos dados de radares e descargas elétricas e, até mesmo, a escolha dos parâmetros para cada um dos algoritmos, bem como o delineamento das métricas de erro adotadas.

Portanto, os primeiros capítulos preocupam-se com a revisão teórica e exploração dos materiais e métodos necessários a construção dos modelos que se propõe. Com isso, o Capítulo 5 explora o desempenho dos algoritmos aplicados ao problema de pesquisa.

Isso se dá pela avaliação dos erros obtidos para cada um deles, em análise as técnicas por grupo, em concorrência geral e comparados ao TITAN. Ainda é explorada a previsão do tamanho da tempestade, avaliada para os eixos maior e menor e a área da elipse.

Por fim, no Capítulo 6 é realizada a discussão das contribuições dos modelos desenvolvidos na pesquisa, bem como alguns estudos que podem sem considerados para trabalhos futuros.

2 CONCEITOS METEOROLÓGICOS

A meteorologia é uma ciência amplamente complexa, uma vez que se dedica ao estudo dos fenômenos atmosféricos. A atmosfera, além de ser extensa, é variável e suscetível a ocorrência de diversos fenômenos que podem afetar a vida humana e a economia (INMET, 2017).

O Brasil "é um país com um grande número de eventos intensos e extremos"(PESSOA, 2014), como tempestades severas. Devido ao potencial desses eventos em causar danos, como deslizamentos, inundações e outros, a previsão do deslocamento de tempestades severas "é um tema atual de pesquisa em Meteorologia"(PESSOA et al., 2012).

Nesse contexto, os conceitos meteorológicos necessários a pesquisa são abordados em quatro partes: A primeira consiste na revisão do processo de acompanhamento de tempestades. A segunda e terceira se dedicam a breve revisão dos dados que permeiam a pesquisa: radares meteorológicos e descargas elétricas atmosféricas. Por último, o software utilizado como base nesta pesquisa, o TITAN, é apresentado, com uma breve revisão dos seus processos.

2.1 ACOMPANHAMENTO DE TEMPESTADES SEVERAS

O acompanhamento de tempestades severas é importante, devido a se tratarem de sistemas precipitantes que podem causar diversos danos, como enchentes e outros relacionados a descargas elétricas atmosféricas. Visto que a agroindústria, setor vulnerável a precipitação e eventos relacionados, é uma atividade econômica predominante no Brasil, informações como a precipitação são relevantes ao desenvolvimento sócio-econômico do país. Nesse sentido, o *nowcasting*, ou previsão a curtíssimo prazo de tempestades, tem grande importância na tomada de decisões (CALHEIROS, 2008).

Um método amplamente aplicado na previsão imediata de tempestades é o de centróide, juntamente com dados de radares meteorológicos. As primeiras pesquisas dessa abordagem se deram por Wilk e Gray (1970), que propuseram a identificação de células convectivas, seu rastreio e previsão pela extrapolação do centróide das células de tempestades.

Mais tarde, Dixon e Wiener (1993) "desenvolveram um robusto sistema de rastreamento e análise de células em tempo real"(WILSON et al., 1998). Os autores mostram que a previsão é realizada a partir de um ajuste linear, obtida com base no his-

tórico das tempestades, ou seja, suas tendências passadas. Além disso, incorporaram o tratamento de fusões e divisões das células com o uso de algoritmos geométricos. Essa metodologia é apresentada na seção 2.4.

Outras abordagens que adotam o algoritmo do tipo centróide são o SCIT (*Storm Cell Identification and Tracking*) apresentado por Johnson et al. (1998) e o TRT (*Thunderstorms Radar Tracking*) por Hering et al. (2004). O primeiro abrange o movimento das tempestades, por meio de dados de radares em três dimensões e é adotado em um sistema mais amplo, o Sistema de Suporte a Decisões de Aviso (WDSS), usado na Austrália. Já o segundo, é um sistema de previsão imediata de tempestades, que utiliza imagens compostas de radares que, a partir de uma sequência de dados, estima o movimento futuro do centróide das tempestades (DESLANDES; RICHTER; BANNISTER, 2008; QUEIROZ, 2009).

A técnica de Previsão e Rastreamento da Evolução de Clusters de Nuvens, denominada ForTraCC, apreentada por Vila et al. (2008), em contrapartida às demais, utiliza-se de imagens de infravermelho de satélites. Seu algoritmo de previsão baseia-se na evolução dos eventos nos períodos de tempo anteriores. Uma ampliação foi proposta por Calheiros (2008), o HydroTrack, que tem por base o ForTraCC e um modelo de predição de precipitação, o Hydro-Estimador, que utiliza de dados de satélites geoestacionários, para uma melhoria na previsão da propagação de sistemas precipitantes.

Outras propostas surgiram a partir do ForTraCC, visando a adaptação para o uso de radares meteorológicos. Queiroz (2009) analisou o uso de radares, com base nos parâmetros VIL, DVIL (desndidade de VIL), altura de máxima refletividade, etc. Analogamente, Bonato (2014) propôs a adaptação para dados de refletividade máxima do radar.

Uma abordagem um pouco diferente das apresentadas foi realizada por Kleina, Matioli e Alvim Leite (2016), que desenvolveram um sistema de identificação, monitoramento e previsão de tempestades elétricas, o qual utilizou de dados de raios, propondo a extrapolação dos dados a curtíssimo prazo.

Nesse contexto, a ciência de acompanhamento de tempestade é relativamente nova e, ainda, não se conhecem medidas diretas da relação dos atributos de uma tempestade (SOS-CHUVA, 2015).

As maiores dificuldades no processo de tomadas de decisões estão associadas à falta de conhecimento sobre os processos no interior das nuvens, em eventos de tempo severo. O uso de dados meteorológicos, sejam coletados por sensores ou modelados, aliados a técnicas computacionais "que possibilitem assimilar inteligentemente esses volumes de dados", permitem "realizar previsões alternativas de tempo"(PESSOA

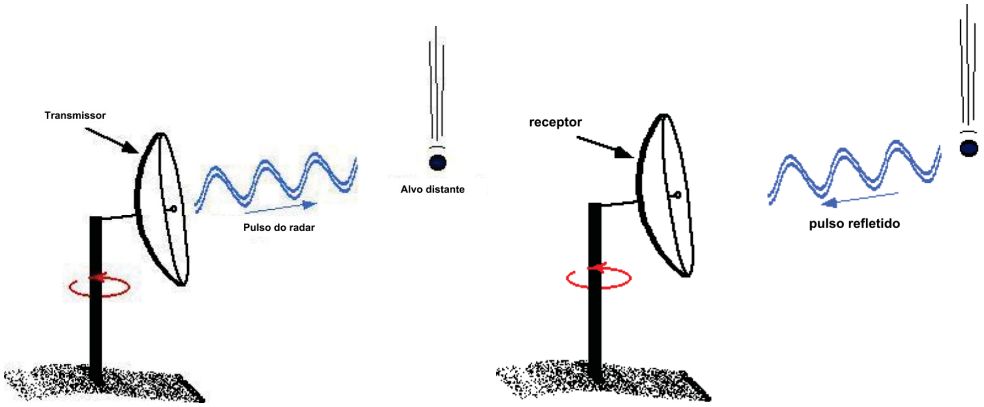
et al., 2012).

Dessa forma, são apresentados dois tipos de dados meteorológicos coletados e que estão presentes nas tempestades. O primeiro é relativo a precipitação e obtido por meio de radares meteorológicos, conforme definido na seção 2.2. O segundo é de dados de descargas elétricas atmosféricas, descritos na seção 2.3.

2.2 RADAR METEOROLÓGICO

O Radar (RAdio Detection And Ranging), apesar de ter sido desenvolvido inicialmente por objetivos militares, tornou-se uma importante ferramenta para a meteorologia. O instrumento, basicamente, consiste na emissão de sinal em frequência de microondas e recebe o eco da reflexão dos alvos, ou objetos (FABRY, 2015), como mostra a FIGURA 1.

FIGURA 1 – FUNCIONAMENTO DO RADAR



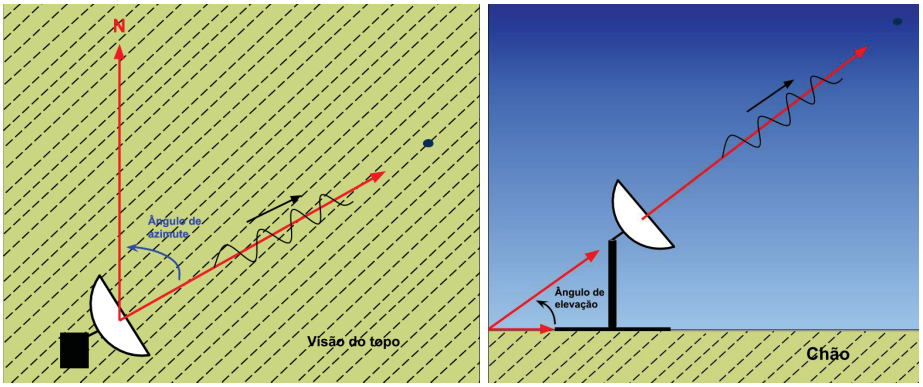
(a) Emissão da onda (b) Recebimento do eco de reflexão do alvo

FONTE: Adaptado de Illinois (2010)

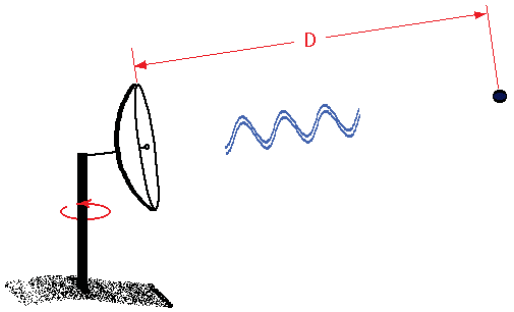
Os radares meteorológicos têm se destacado na vigilância da atmosfera, visto a capacidade que as microondas possuem de penetrar áreas de precipitação (TEIXEIRA, 2010).

A coleta de dados de um radar, em que a localização de um alvo é dada por três informações, é apresentada na FIGURA 2. A primeira (a) se refere a rotação em relação ao seu prório eixo, ao norte geográfico, chamamos ângulo de azimute da varredura. A segunda (b) é a elevação, que se trata do ângulo de rotação vertical do instrumento, ou seja, em relação ao solo. A terceira e última (c), se refere à distância da leitura ao radar.

FIGURA 2 – LOCALIZAÇÃO DO ALVO



(a) Azimute (b) Elevação

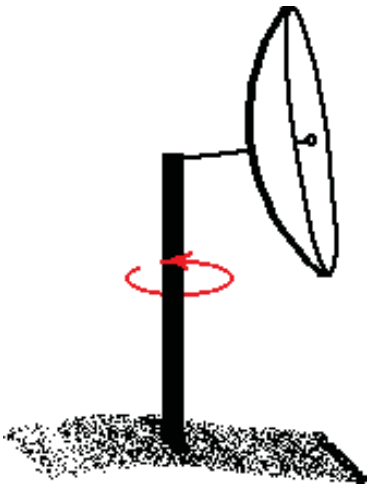


(c) Range

FONTE: Adapatado de Illinois (2010)

Os dados coletados podem ser projetados em um plano, fixando um ângulo de elevação. Quando fixa-se o ângulo de elevação e o ângulo de azimute é variado (procedimento representado pela FIGURA 3) obtém-se uma visão horizontal dos dados que é denominada PPI (Plan Position Indicator) (DAMIAN, 2012).

FIGURA 3 – EXEMPLO DA LEITURA PPI



FONTE: Illinois (2010)

O PPI pode ser obtido diretamente da estratégia de varredura dos dados, ou

da redução de um dado volumétrico para sua projeção no plano. Trata-se "forma básica de visualização das variáveis medidas pelo radar"(OLIVEIRA, 2014).

Uma das variáveis coletadas pelo radar é a refletividade (Z), que pode ser dada pela equação 2.1:

$$z = c_r r^2 p_r \tag{2.1}$$

onde p_r é a energia refletida do radar, r o range, ou distância do radar ao alvo e, por último, c_r é uma constante do radar que agrega informações como o comprimento de onda e de pulso, forma e largura do feixe, potência transmitida, entre outras (DAMIAN, 2012).

Usualmente os fatores de refeltividade são dados em decibéis (dB), de onde tem-se:

$$dBZ = 10log(z) \tag{2.2}$$

O valor de dBZ auxilia a identificar tempestades, como usado na seção 2.4, em que valores mais altos indicam um possível ponto de tempestade ativo.

Um dos produtos gerados a partir dos dados de refletividade é o VIL: conteúdo líquido verticalmente integrado (verticaly integrated liquid), que busca estimar a quantidade de água líquida verticalmente. É útil na identificação de células de tempestade de forma rápida e eficaz e, também, para estimar massa de granizo e potencial da rajada de vento (FABRY, 2015).

O VIL é definido pela equação 2.3:

$$VIL = 3,44 \times 10^6 \int \left(\frac{z_i + z_{i+1}}{2} \right)^{\frac{4}{7}} dh \tag{2.3}$$

Em que h representa a altura dos limites da uma camada dada e z_i e z_{i+1}, respectivamente, representam a refletividade no limite inferior e superior (DAMIAN, 2012). Essa integração ocorre desde o ângulo de elevação mais baixo ao mais alto, em que a água líquida é convertida com base na refletividade e na relação Z-R, em que R é taxa de precipitação. Mais informações sobre essa relação são mostradas por Rinehart (2004).

2.3 DESCARGAS ELÉTRICAS ATMOSFÉRICAS

O fluxo intenso de corrente elétrica de curta duração, ocorrido na atmosfera, é denominado descarga elétrica atmosférica, e pode ou não atingir a superfície terrestre (PESSOA, 2014). Esse fenômeno se inicia pelo acúmulo de cargas elétricas na nuvem, que excedem a capacidade isolante do ar. Com isso, elétrons que estão em uma região de carga negativa, se deslocam para uma região positiva (RINDAT, 2018). A ocorrência de descargas pode se restringir apenas as nuvens ou também em contato com o solo, como apresenta o QUADRO 1, segundo (PESSOA, 2014).

QUADRO 1 – TIPOS DE DESCARGAS ATMOSFÉRICAS QUANTO AO PERCURSO

Tipo	**Definição**
Intra-nuvem (IN)	Ocorre no interior da nuvem.
Nuvem-nuvem (NN)	O fluxo de corrente elétrica se dá do interior de uma nuvem à outra.
Nuvem-solo (NS)	Ocorre entre a nuvem e a superfície terrestre.

FONTE: A autora (2018)

As descargas do tipo NN são as mais comuns, enquanto isso, as de NS representam cerca de 20% das ocorrências (SHIGA, 2007; PESSOA, 2014; RINDAT, 2018).

As descargas do tipo Nuvem Solo podem ser descendentes, quando tem início na nuvem e se move em direção ao solo e ascendentes quando a trajetória é oposta, sendo que esse útimo tipo é menos comum. Podem ser classificadas em positiva ou negativa, com base em seu sinal de polaridade, de acordo com Shiga (2007), conforme o QUADRO 2.

QUADRO 2 – TIPOS DE DESCARGAS ATMOSFÉRICAS QUANTO À POLARIDADE

Termo	**Definição**
Positivo	Quando a nuvem está carregada positivamente, sendo neutralizada por uma descarga em que há um fluxo ascendente de elétrons.
Negativo	Quando a nuvem está carregada negativamente, sendo neutralizada por uma descarga em que há um fluxo descendente de elétrons

FONTE: Shiga (2007)

As descargas que ocorrem entre a nuvem e a superfície terrestre são as de interesse neste trabalho. Devido a sua interação com o solo, são as que podem causar danos materiais ou à vida humana. Além disso, os raios positivos podem ser

dominantes na dissipação das tempestades, apesar de representarem apenas 10% (aproximadamente) das ocorrências NS (HEIDLER et al., 2008).

2.4 O SOFTWARE TITAN

O TITAN (*Thunderstorm Identification, Tracking, Analysis and Nowcasting* - Identificação, rastreamento, análise e previsão de tempestades a curto prazo) é um software que aborda a identificação, o rastreamento, análise e a previsão a curto prazo de tempestades em tempo real, a partir de dados de radares meteorológicos (DIXON; WIENER, 1993).

A sua primeira versão foi desenvolvida em 1980 e passou a ser mantida e aprimorada pelo Centro Nacional de Pesquisa Atmosférica (NCAR). Atualmente, se encontra na versão TITAN 5, é livre e possui um sistema bem completo, desde a identificação até a visualização de tempestades (NCAR, 2016). Tem sido utilizado em todo o mundo, destacando-se a Austrália, África do Sul e Ásia (HAN et al., 2009). No Brasil, o SIMEPAR e o IPMet (GOMES; HELD, 2006) são exemplos de orgãos que o utilizam.

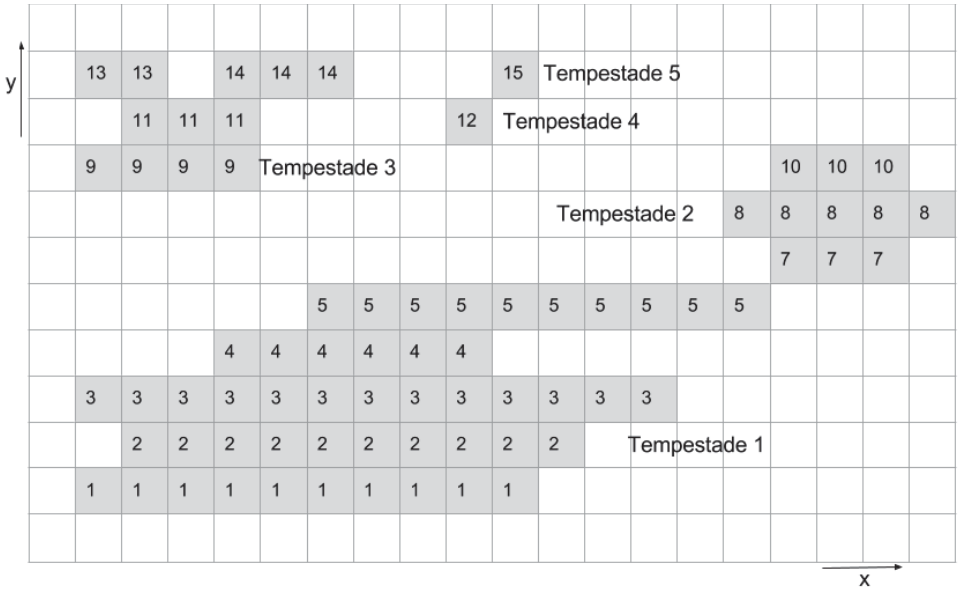
O software realiza o acompanhamento de tempestades gerando algumas características sobre o evento que permitem realizar outros estudos sobre as tempestades identificadas. Assim, cada um dos três processos estão descritos a seguir.

2.4.1 Identificação de tempestades

Uma vez que, "uma tempestade é definida como uma região contígua excedendo limites de refletividade e tamanho"(DIXON; WIENER, 1993), o TITAN realiza o processo de identificação admitindo-se limitantes para os valores de tamanho e refletividade observados de tempestades, obtidos por meio de dados de radares meteorológicos.

De acordo com Bonato (2014), esse processo se dá em 2 passos: no primeiro "identifica-se sequências de pontos contíguos (runs) em uma determinada direção, nos quais os valores de refletividade superam um determinado limiar"; enquanto no segundo "agrupa-se runs adjacentes, formando as tempestades". Esse processo é exemplificado na FIGURA 4, que representa os dados de refletividade que estão acima do limiar estabelecido em uma grade cartesiana, em que a leitura dos dados é feita no sentido crescente nos eixos x e y, respectivamente.

FIGURA 4 – EXEMPLO DO PROCESSO DE IDENTIFICAÇÃO DO TITAN



FONTE: Adaptadado de Dixon e Wiener (1993)

Em seguida são verificadas as regiões adjacentes, horizontalmente são identificadas na figura como um mesmo número, fazendo-se a verificação da adjacência vertical obtém as tempestades 1 a 5. Aplicando-se o limiar de tamanho, as duas últimas tempestades são descartadas. Para o caso tridimensional, de acordo com Bonato (2014), isso ocorre identificando pontos adjacentes não apenas na direção x e y, mas também levando em consideração o eixo z.

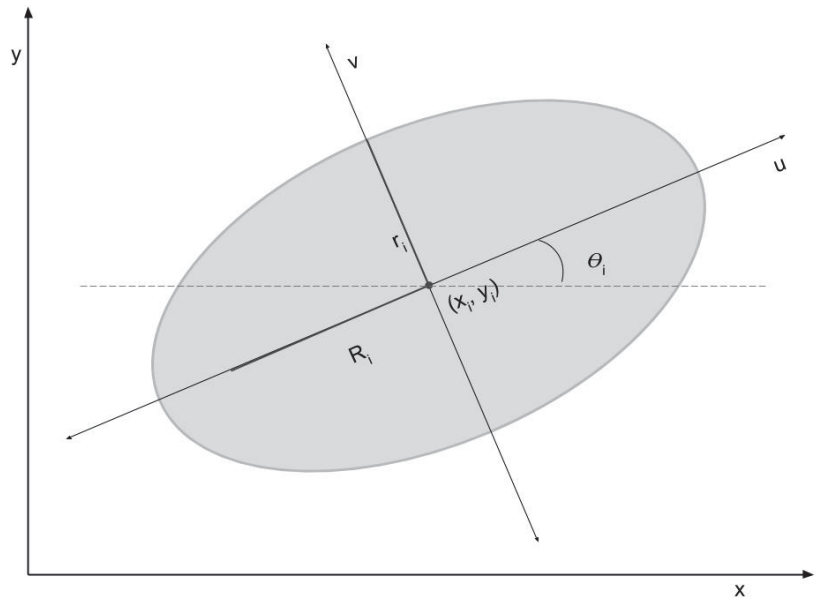
Para cada tempestade identificada são geradas algumas características que podem ser observadas pelos radares meteorológicos ou calculadas a partir desses dados e com base em seus conceitos físicos. Elas podem se referir a posição, ao deslocamento e a intensidade da tempestade.

O evento pode ser representado na forma de polígono ou elipse, cuja exemplificação pode ser visualizada na FIGURA 5. Em que o par (x_i, y_i) representa o centróide ou centro de massa da tempestade i, enquanto θ_i a orientação da elipse e, por fim, R_i e r_i os seus semi-eixos maior e menor, repectivamente.

Para o caso bidimensional, considerando-se a representação das células em elipse, assim como estabelecido nessa pesquisa, as características geradas são listadas no QUADRO 3.

Para o caso tridimensional, ao invés de apresentar a área da tempestade são apresentadas a área projetada e a área média, resultantes das áreas de cada volume. As demais características são mantidas e ainda são acrescentadas informações, como o volume, a altura da célula, a altura do valor máximo de refletividade, entre outras.

FIGURA 5 – EXEMPLO DA REPRESENTAÇÃO DA TEMPESTADE EM ELIPSE



FONTE: Adaptadado de Dixon e Wiener (1993)

QUADRO 3 – DESCRIÇÃO DAS CARACTERÍSTICAS DAS TEMPESTADES

Características	**Descrição**
Centróide	Latitude e Longitude referente a posição do centro da elipse.
Orientação	Ângulo de rotação da elipse em relação ao Norte.
Velocidade	Velocidade calculada do deslocamento.
Direção	Direção do deslocamento.
Eixo maior	Tamanho, em quilômetros, do eixo maior da elipse.
Eixo menor	Tamanho, em quilômetros, do eixo menor da elipse.
Área	Área calculada da tempestade.
VIL	Conteúdo líquido verticalmente integrado.
dBZ	Representa o valor de refletividade em escala logarítmica.
Massa de granizo	Estimativa da massa de granizo na tempestade.

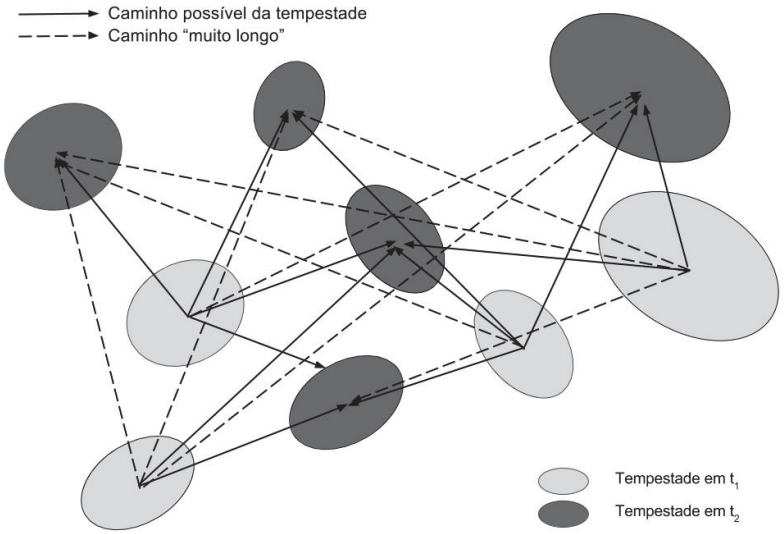
FONTE: A autora (2018)

2.4.2 Rastreamento de tempestades

O rastreamento é realizado com base nas características de pontos de tempestades já identificados. Esse processo consiste, genericamente, em "determinar o movimento correspondente entre células de tempestades em imagens de radares sucessivas"(HAN et al., 2009). Com isso, é possível obter um histórico de deslocamento para cada tempestade conhecida.

Para determinar esse movimento, o TITAN usa a técnica de sobreposição que busca identificar formas correspondentes de células de tempestades em dois períodos de tempo sucessivos. Em seu algoritmo ainda implementa um limiar de velocidade máxima de deslocamento (HAN et al., 2009). O processo realizado pelo software é exemplificado na FIGURA 6.

FIGURA 6 – EXEMPLO DO PROCESSO DE RASTREAMENTO DO TITAN



FONTE: Adaptadado de Dixon e Wiener (1993)

Segundo Dixon e Wiener (1993), alguns pressupostos intuitivos podem ser feitos quanto o melhor conjunto de rastreio:

1. Incluir os menores caminhos (deslocamentos);
2. Juntar tempestades que são semelhantes entre si, em relação ao seu tamanho, formato, etc; e

3. Não exceder o limite de velocidade máximo assumido, em que uma tempestade pode se deslocar em um período Δt (no exemplo isso é representado como caminhos denominados muito longos).

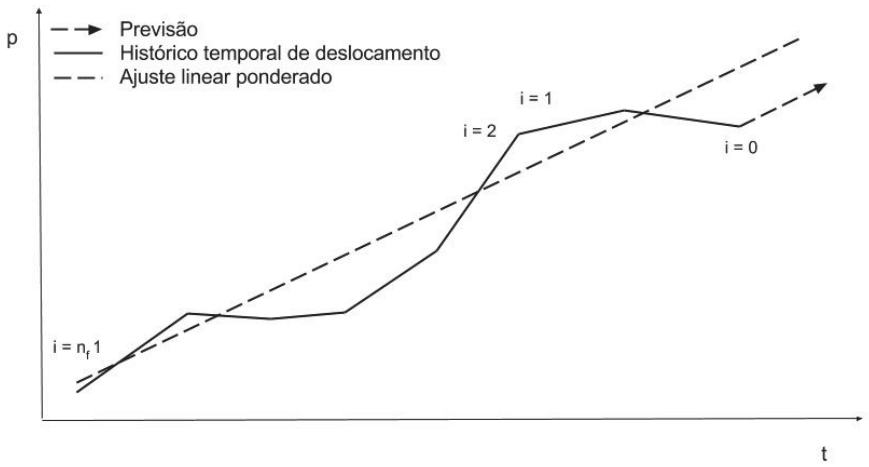
O limitante máximo de velocidade pode, de acordo com Han et al. (2009), causar erros no rastreamento, uma vez que a velocidade é calculada em relação ao centróide ponderado pela refletividade. Como rapidamente o formato de uma célula de tempestade pode mudar, isso causaria aleatoriedade indesejada no deslocamento do centróide, reduzindo a confiabilidade do rastreamento.

2.4.3 Previsão a curto prazo de tempestades

A construção de um histórico para esse tipo de evento permite a previsão do seu deslocamento para até 60 minutos, sendo cada período de tempo 10 minutos, ou seja, permite a previsão de suas seis posições seguintes. Isso é realizado pela técnica a partir da identificação de uma tempestade em pelo menos três períodos subsequentes de tempo.

Esse processo se dá pela extrapolação linear e com um coeficiente de ajuste que tem como base o histórico das características da tempestade, um exemplo desse ajuste pode ser visualizado na FIGURA 7.

FIGURA 7 – EXEMPLO DO PROCESSO DE PREVISÃO DO TITAN



FONTE: Adaptado de NCAR (2016)

A extrapolação da tempestade é feita utilizando as seguintes propriedades: Área projetada do centróide, centróide volumétrico (Z), refletividade do centróide, altura,

máximo valor de dBZ, fluxo de precipitação, massa, velocidade, direção, área projetada e volume (DIXON; WIENER, 1993; NCAR, 2016).

O TITAN é utilizado em dois processos deste trabalho: O primeiro é na geração do histórico das tempestades, ou seja, na identificação e rastreamento. O segundo é no fornecimento de previsões para esses eventos, a fim de possibilitar avaliação do desempenho das abordagens propostas.

Dessa forma, apresentou-se os conceitos meteorológicos que permeiam este este trabalho. Os dados de radares meteorólogícos aparecem no processamento do TITAN, enquanto os de descargas atmosféricas necessitam ser incorporados nesses dados, como é discutido no Capítulo 4.

Com o objetivo de estudar as técnicas de Aprendizado de Máquina para a previsão do deslocamento de tempestade, como foi discutido nesse capítulo, no Capítulo 3 é dissertado a respeito dessa área do conhecimento e dos métodos propostos.

3 APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina (AM) é uma subárea de Inteligência Artificial (RÄTSCH, 2004) que, de acordo com Libralão et al. (2003), "estuda o desenvolvimento de métodos capazes de extrair conceitos (conhecimento) a partir de amostras de dados", ou seja, que permitam que a máquina obtenha aprendizado.

Diversas áreas do conhecimento utilizam esse método. Segundo Harrington (2012), o campo de AM "está na interseção das áreas de informática, engenharia, estatística e, muitas vezes, aparece em outras disciplinas". Visto que trata-se de uma área do conhecimento ampla, nesse capítulo são abordados os principais conceitos da área, o tipo de algoritmo para os objetivos que a pesquisa se propõe, bem como cada um dos métodos que a constitui.

Novas aplicações de algoritmos de AM são desenvolvidas todos os dias em diferentes tipos de problemas. Nesse contexto, Mohri, Rostamizadeh e Talwalkar (2012) descrevem as maiores classes de problemas de Aprendizado, conforme apresentado no QUADRO 4.

QUADRO 4 – PRINCIPAIS CLASSES DE PROBLEMAS DE APRENDIZADO

Problema	**Definição**
Classificação	Atribuição de uma categoria para cada item do conjunto de dados.
Regressão	Predição de um valor real para cada item do conjunto de dados.
Ranking	Ordenação dos item do conjunto de dados segundo um mesmo critério.
Clusterização	Agrupamentos de itens semelhantes entre si numa mesma região ou grupo.
Redução de dimensionalidade	Transformar a representação inicial dos dados em uma dimensão menor preservando suas propriedades.

FONTE: A autora (2018)

Pela diversidade desses problemas e de acordo com Mohri, Rostamizadeh e Talwalkar (2012), qualquer campo que precise interpretar e atuar em dados pode se beneficiar dessas técnicas.

Diversos estudos de meteorologia apresentam abordagens por aprendizado de máquina, principalmente os problemas de Classificação, Regressão e Clusterização. Alguns exemplos são na previsão de clima (PESSOA, 2004), previsão de vazões naturais afluentes (GUILHON; ROCHA; MOREIRA, 2007), estudo de padrões climáticos

sazonais (ANOCHI; SILVA, 2009), previsão probabilística de alagamentos no Município de Curitiba (LOHMANN, 2011), estimativa de chuva (SANTOS, 2014), previsão climática de precipitação (ANOCHI; CAMPOS VELHO, 2016), identificação de eventos de tempo severos (SILVA, 2017), entre outros.

Embora a meteorologia seja a grande área envolvida nesses exemplos, existem aplicações bem variadas, como clima, vazão, precipitação, etc. O uso dessa técnica na área pode ser atribuído a quantidade de dados e, por muitas vezes, as características dos fenômenos não serem exatas ou por possuirem padrão desconhecido.

Dada a natureza dos dados que se dispõe e do problema de pesquisa, a aplicação é um problema de regressão, descrito na seção 3.2.

Optou-se, neste trabalho, por métodos de AM, pois apresentam grande potencial em aprender características de dados e seus relacionamentos. Além disso, a execução de novas entradas de dados ocorre de forma muito rápida, uma vez que um modelo é treinado baseando-se nesse tipo de abordagem e com uma taxa de confiabilidade desejada.

3.1 DEFINIÇÕES

Ao trabalhar com Aprendizado de Máquina é importante conhecer algumas expressões que são usadas com frequência e se referem a conceitos relacionados ao processo de aprendizado, aos dados utilizados e de sua partição em conjuntos.

Os principais termos utilizados para se refereir aos dados são exemplos, atributos ou características e rótulos, como descrito no QUADRO 5 de acordo com Mohri, Rostamizadeh e Talwalkar (2012).

QUADRO 5 – TERMOS UTILIZADOS EM REFERÊNCIA AOS DADOS

Termo	**Definição**
Exemplos	Instâncias de dados utilizadas no processo de Aprendizado.
Atributos	Conjunto de atributos associado a um exemplo, normalmente representado por um vetor.
Valores esperados	Valores ou categorias associadas a cada exemplo.

FONTE: A autora (2018)

Com base nessas informações, é possivel dizer que atributos são as características de cada item a ser aprendido, enquanto os rótulos representam o valor esperado

do aprendizado. Em outras palavras, um exemplo consiste em um vetor de atributos juntamente como seu valor esperado.

Existem diferentes tipos de aprendizagem de máquina que, segundo Mohri, Rostamizadeh e Talwalkar (2012), se diferenciam pelo tipo de dados disponíveis para o processo de aprendizado, o método e a ordem em que esses dados são recebidos e, por fim, como os dados são utilizados na avaliação do algoritmo. Dois desses tipos são apresentados no QUADRO 6.

QUADRO 6 – TIPOS DE APRENDIZAGEM DE MÁQUINA

Tipo	**Descrição**	**Principais exemplos**
Supervisionado	No processo de aprendizado são apresentados atributos e seus rótulos esperados, para que o modelo possa ser ajustado de acordo com as saídas desejadas.	Problemas de classificação e Regressão.
Não supervisionado	Não é conhecido qualquer rótulo para os atributos, de modo que a aprendizagem deve ser realizada apenas com base nos atributos apresentados.	Problemas de Agrupamento e Redução de dimensionalidade.

FONTE: A autora (2018)

No processo não supervisionado, como a rede desconhece os valores esperados, o aprendizado baseia-se apenas nos dados de entrada. Esse tipo de modelo pode ser difícil de avaliar quantitativamente, pois, muitas vezes, não são conhecidos os valores esperados (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012).

No caso supervisionado, se conhece os valores esperados e espera-se que com os exemplos apresentados ao método, ele encontre "uma função determinística que mapeie qualquer entrada para um resultado, de modo que o desacordo com futuras observações seja minimizado"(HERBRICH, 2001). Em outras palavras, são apresentados os valores desejados e o objetivo é aprender uma regra geral que seja capaz de mapear as entradas para tais saídas. Quando isso ocorre, indica que o modelo apresenta a capacidade de generalização.

A generalização de um modelo supervisionado é importante, pois mostra a sua capacidade em aprender as características dos dados trabalhados e indica que está apto a prever novas observações feitas do mesmo tipo de dado.

Os dados utilizados no processo de AM supervisionado são divididos em, pelo menos, conjunto de treinamento e teste, conforme descritos no QUADRO 7 de acordo com Mohri, Rostamizadeh e Talwalkar (2012). Em geral, o conjunto de treinamento

consiste de 60% a 90% dos dados.

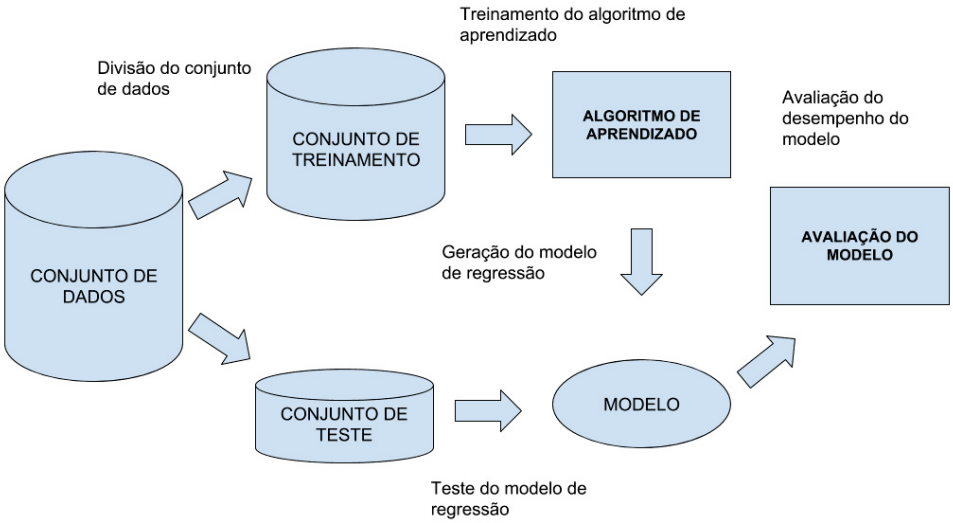
QUADRO 7 – DIVISÃO DO CONJUNTO DE DADOS NAS APLICAÇÃO DE AM

Conjunto	**Descrição**
Treinamento	Conjunto de exemplos utilizados no processo de aprendizagem do algoritmo.
Teste	Conjunto de exemplos utilizados para avaliar (testar) o modelo treinado.

FONTE: A autora (2018)

Nesse contexto, a aprendizagem de máquina ocorre a partir dos exemplos do conjunto de treinamento. Quando se tem o modelo de Aprendizado treinado, então são apresentados os atributos dos dados do conjunto de teste, sem seus respectivos rótulos. Com isso, pode-se ter uma ideia da precisão do algoritmo comparando a predição dos dados de teste e os valores esperados, como descreve Harrington (2012), ilustrado na FIGURA 8.

FIGURA 8 – FLUXOGRAMA DO PROCESSO DE APRENDIZADO SUPERVISIONADO



FONTE: A autora (2018)

3.2 REGRESSÃO

Regressão é uma das classes dos problemas em que AM se aplica e, como apresentado no QUADRO 4, trata-se da "previsão de um valor numérico"(HARRINGTON, 2012). Como esse tipo de algoritmo é treinado apresentando-se o valor de previsão desejado, se classifica como Aprendizado supervisionado.

Nesse tipo de problema, espera-se encontrar uma função $f(x)$ que descreva a relação entre x e y, caso exista, como dado pela equação 3.1:

$$y = f(x) + \epsilon, \tag{3.1}$$

onde y representa os rótulos, ou seja, os valores que se deseja prever e x a váriavel regressora, da forma $[x_1, x_2, ..., x_k]$, de modo que cada x_j representa o j-ésimo exemplo de aprendizagem, com $j = 1, ..., k$, e k é o número de características de entrada. Desse modo, as variáveis "são usadas principalmente para prever ou explicar o comportamento de y"(SEBER; WILD, 2003) e ϵ representa a diferença entre o valor estimado pela função e o rótulo desejado.

A função $f(x)$ é denominada função regressora e permite a previsão do valor aproximado de y, dado os atributos x. Assim, a regressão pode ser classificada como linear e não linear, se $f(x)$ for uma função do tipo linear ou não.

Existem vários grupos de métodos que podem ser aplicados em problemas de regressão, como o *Ensemble* e Modelo Linear que são utilizados nesse trabalho e descritos nas seções 3.3 e 3.4, respectivamente.

3.3 MÉTODOS DE ENSEMBLE

Em um método de regressão simples geram-se hipóteses a partir do conjunto de dados e seleciona-se a hipótese que apresenta melhor consistência aos dados, como os métodos de Modelo Linear, por exemplo. Nos métodos de *ensemble* a função regressora é determinada pela combinação das hipóteses geradas, ou seja, trata-se da combinação da predição de vários regressores, com alguma variação nos parâmetros ou no seu conjunto de aprendizado (COENEN; PREECE; MACINTOSH, 2011).

Em problemas de regressão, geralmente a combinação das predições de cada um dos membros do conjunto é realizada por média simples, mas também pode ocorrer pela ponderação dos preditores individuais.

Segundo Brilhadori e Lauretto (2013), *ensemble* é de "um paradigma de aprendizado em que um grupo finito de propostas alternativas é utilizado para a solução de um dado problema". Essa combinação de propostas é interessante quando, apesar de distintas, cada uma delas se mostra relevante para o problema.

Esse tipo de abordagem pode melhorar o desempenho do modelo resultante, uma vez que, de acordo com Wichard e Ogorzalek (2004), "o erro de generalização do conjunto é menor que a média do erro de generalização dos membros do conjunto".

Para isso, o desempenho individual de cada preditor deve ser bom e, quando comparado aos demais, ter comportamento distinto (COELHO et al., 2006).

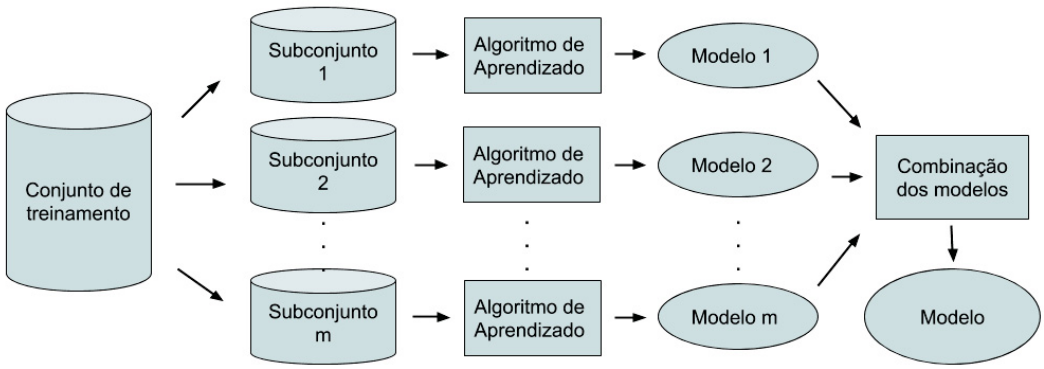
Assim, um *ensemble* consiste na obtenção de preditores altamente precisos a partir da combinação de outros menos precisos. Nesse sentido, Dietterich (2000) afirma que o desempenho e a diversidade dos membros individuais é uma condição necessária e suficiente para que o *ensemble* apresente melhor performance. De fato, a redução do erro de generalização do conjunto está relacionada com o aumento da variância entre os preditores (KROGH; VEDELSBY, 1995; BROWN; WYATT; TIŇO, 2005; AUDHKHASI et al., 2012).

Quanto à diversidade dos membros, ainda é possível comprender que um conjunto de preditores iguais ou muito semelhantes combinados possuem resultados iguais ou muito próximos de qualquer regressor individual, apenas causando um aumento do tempo e custo computacional.

Visto a importância de reduzir a correlação entre os membros do conjunto, na construção do método, existem algumas formas de introduzir a diversidade nesses modelos. Podem ser relacionadas ao conjunto de dados ou ao algoritmo utilizado.

Os métodos mais pesquisados são aqueles que apresentam variação nas amostras do conjunto de treinamento, como apresentado na FIGURA 9. Criam-se, do conjunto de treinamento, um número desejado (m) de amostras de dados que são treinadas em um mesmo algoritmo de aprendizado, gerando-se um modelo para cada um dos subconjuntos.

FIGURA 9 – ENSEMBLE COM VARIAÇÃO NO CONJUNTO DE TREINAMENTO



FONTE: A autora (2018)

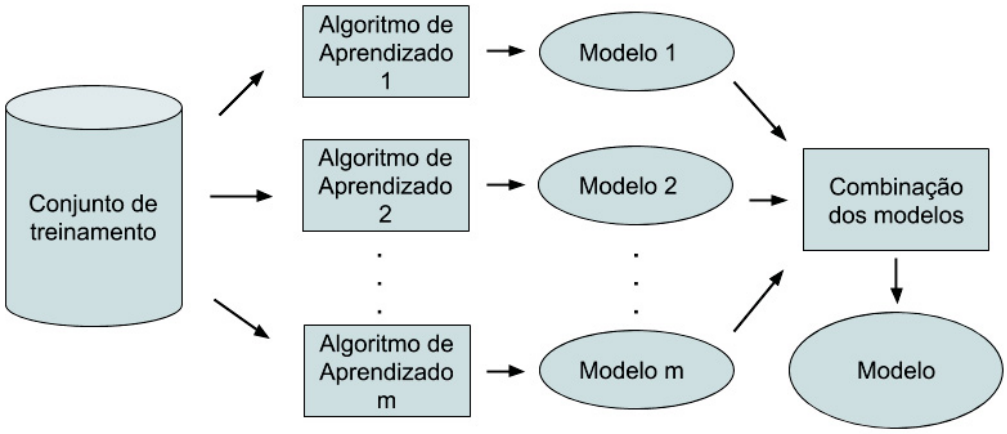
Nesse tipo de método existe a manipulação dos exemplos de treinamento e, por isso, um melhor desempenho é encontrado para algoritmos instáveis, pois esse tipo de algoritmo é mais sensível ao conjunto de dados. Essa sensibilidade é devido ao fato de pequenas alterações na amostra de dados resultarem em mudanças significativas no

modelo. São exemplos de técnicas instáveis as redes neurais e as árvores de decisões (DIETTERICH, 2000).

As técnicas de reamostragem permitem a variação no fornecimento do conjunto de dados para o treinamento. Elas podem aparecer gerando diferentes subconjuntos da amostra de dados, mantendo todos os atributos do conjunto original ou, ainda, formando subconjuntos diferentes a partir deles (COELHO et al., 2006). São exemplos de algoritmos que utilizam esse tipo de técnica: *Bagging* e *Boosting*.

Outra metodologia encontrada na literatura para garantir a diversidade dos modelos é realizar o treinamento dos dados por m variações de algoritmos de aprendizado, resultando em um modelo para cada um deles, como apresentado na FIGURA 10.

FIGURA 10 – ENSEMBLE COM VARIAÇÃO NOS ALGORITMOS DE APRENDIZAGEM



FONTE: A autora (2018)

Nesse tipo de abordagem a diversificação dos algoritmos pode se dar pela manipulação da arquitetura ou mudanças de parâmetros, que é quando se utiliza o mesmo tipo de aprendizado, com diferentes configurações, como por exemplo, mudando-se alguns parâmetros do algoritmo ou da sua inicialização.

Outra forma é a construção de conjuntos, ou *ensembles*, heterogêneos, que é o uso de diferentes métodos, como várias topologias de redes neurais, ou mesclar algoritmos bem diferentes. O problema é que nesse tipo de conjunto não há garantia da melhoria no erro de generalização, uma vez que "o uso de diferentes paradigmas leva a componentes com diferentes especialidades e precisões, que podem apresentar diferentes desempenhos e, com isso, diferentes padrões de generalização"(COELHO et al., 2006).

Nesse contexto, os *ensembles* também podem ser chamadas de métodos de randomização, pois introduzem ou exploram a aleatoriedade no algoritmo de apren-

dizado. Assim, se obtém um conjunto com preditores individuais "mais ou menos fortemente diversificados"(GEURTS; ERNST; WEHENKEL, 2006).

As técnicas que foram utilizadas neste trabalho são *Bagging*, *Random Forest* e *Extra Trees*, apresentadas nas subsessões 3.3.2 a 3.3.4. Em que todas utilizam do algoritmo de árvore de decisão (subseção 3.3.1) que, de acordo Koronacki, Ras e Wierzchon (2009), em comparação a outras técnicas de Aprendizado de Máquina, apresenta um rápido processamento na fase de treinamento do modelo.

3.3.1 Árvore de decisão

O algoritmo de árvore de decisão CART (*Classification and Regression Trees*) foi desenvolvido por Breiman et al. (1984). Essa abordagem consiste da subdivisão sucessiva da amostra de dados, até que não seja mais possível dividí-la, ou todos os seus valores desejados sejam iguais.

O particionamento dos dados ocorre de forma recursiva e binária, ou seja, em cada partição escolhe-se um atributo do conjunto de dados e um valor, em que todos os exemplos que possuem o atributo da divisão maior que o valor estabelecido formam a partição direita dos dados, enquanto, os demais, a da esquerda.

Esse tipo de método é interessante quando se tem dificuldade na construção de um modelo global, devido a interação dos atributos ocorrerem de forma complexa. Nos problemas reais existem muitas não-linearidades, dificultando o ajuste de um único modelo global Harrington (2012).

O método é aplicável tanto ao problema de regressão, quanto ao problema de classificação. Devido ao contexto desta pesquisa, a discussão de árvores CART se restringe à regressão.

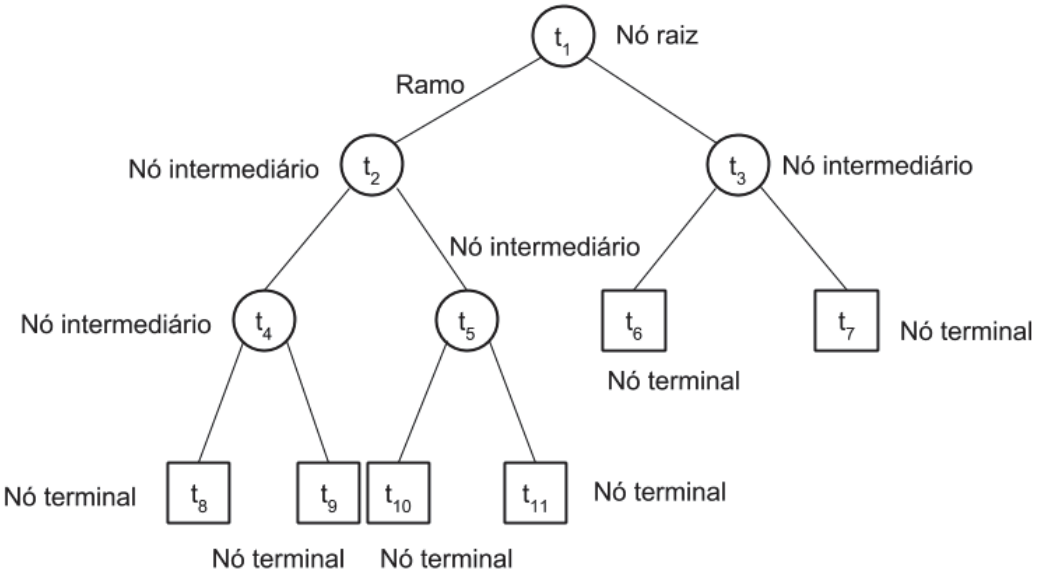
Segundo Breiman et al. (1984), para se obter um preditor em árvore é necessário estabelecer critérios para os três objetivos listados a seguir:

1. Selecionar a divisão de cada nó intermediário;
2. Estabelecer um nó como terminal ou folha; e
3. Atribuir o valor estimado de y em cada folha.

Nesse sentido, nó intermediário ou decisório é aquele que possui uma ramificação subsequente e consiste numa decisão, ou seja, divide a amostra em duas novas partições, gera nós filhos. Nesse sentido, o nó terminal ou folha é o nó que não precede

mais partições e deve levar ao valor estimado de y, conforme pode ser verificado na FIGURA 11, em que cada nó é representado por t.

FIGURA 11 – ESTRUTURA DE UMA ÁRVORE DE DECISÃO BINÁRIA



FONTE: A autora (2018)

Para determinar cada nó, a melhor divisão dentro do espaço de busca de atributos e valores de treinamento é escolhida. Como o "melhor"é algo muito subjetivo, é preciso estabelecer um critério de impureza que permita mensurar a qualidade da divisão de cada nó.

Em árvores de regressão CART, os valores esperados são estimados pela média obtida dos valores observados do conjunto de treinamento, de acordo com o subconjunto, ou nó, a que o dado pertence. Tomando $\hat{y}_t$ como o valor estimado no nó, em que Y_t é o conjunto dos n valores desejados, dos exemplos pertencentes ao nó t, tem-se:

$$\hat{y}_t = \frac{1}{n} \sum_{i=1}^{n} Y_{ti} \tag{3.2}$$

Existem diversas métricas que podem ser adotadas como critério de impureza, o críterio dos mínimos quadrados é bastante utilizado em problemas de regressão em geral e, até mesmo, em outros tipos de árvores de decisões. Entretando, para

CART, Breiman et al. (1984) propõe o uso da média do erro absoluto. Assim, é possível calcular o resíduo em cada nó t, pela equação 3.3:

$$R(Y_t) = \frac{1}{n}\sum_{i=1}^{n} \mid Y_{ti} - \hat{y}_t \mid \tag{3.3}$$

O valor estimado para o nó t, representado por $\hat{y}_t$, trata de uma constante para cada nó da árvore, como pode ser observado pela equação 3.2 que o define.

Dessa forma, são estabelecidos a regra para atribuição do valor estimado e o critério de impureza, equação 3.2 e 3.3, respectivamente, ainda é necessário definir as regras usadas para tornar um nó intermediário em terminal. Para isso, com base em Breiman et al. (1984), três critérios são seguidos:

1. Todos valores desejados no nó possuem o mesmo valor;
2. Todas as novas divisões do conjunto, que minimizam a equação 3.3, geram pelo menos um nó com um número de elementos menor do que o mínimo estabelecido; e
3. A variação da impureza dos dois nós filhos gerados, melhoram a do nó pai muito pouco, menos que um limiar estabelecido.

Com base nesses três critérios são alcançados todos os elementos que foram apresentados como necessários ao crescimento de uma árvore de decisão. O processo de seleção pelo melhor par de nós filhos ou, caso um dos critérios sejam alcançados, a formação do nó terminal é apresentado no ALGORITMO 1, com base em Harrington (2012).

Para a execução do algoritmo, são necessários três parâmetros, $tolS$ que é o valor mínimo de melhora que o algoritmo deve alcançar para dividir em dois nós filhos e $tolN$ que é o número mínimo de exemplos admitidos em cada nó. Por fim, tem-se a matriz D, que consiste de todos os exemplos do nó, representa uma amostra com n exemplos e k atributos:

$$D = \begin{pmatrix} x_{11} & \dots & x_{1k} & y_1 \\ x_{21} & \dots & x_{2k} & y_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nk} & y_n \end{pmatrix} \tag{3.4}$$

O processo de crescimento de árvores é iterativo, no qual executa-se a busca pela melhor divisão de cada nó até que todos os ramos terminem em uma folha.

ALGORITMO 1 – ÁRVORES DE DECISÃO (SELEÇÃO DA MELHOR DIVISÃO DO NÓ)

Parâmetros: $tolS$, $tolN$ e D
Início
1 Inicialize $s = R(D_{\cdot k+1})$ e $a = None$
2 Tome n o número de linhas da matriz D
3 Tome $k+1$ o número de colunas da matriz D
4 **Para** $p = 1$ até k **faça**:
5 **Para** $m = 1$ até n **faça**:
6 Tome $D_e = \{x_i \mid x_{ip} > x_{ij}|, \forall i = 1, ..., n\}$
7 Tome $D_d = \{x_i \mid x_{ip} \leq x_i, \forall i = 1, ..., n\}$
8 **Se** $(\#D_{(e)} \geq tolN)$ ou $(\#D_{(d)} \geq tolN)$ **então**:
9 Determine $s_{novo} = R(D_{(e)\cdot k+1}) + R(D_{(d)\cdot k+1})$
10 **Se** $s_{novo} < s$ **então**:
11 Atualize $s = s_{novo}$
12 Tome $a = p$
13 Tome $v = x_{pm}$
14 **Fim se**
15 **Fim se**
16 **Fim para**
17 **Fim para**
18 **Se** $((R(D_{\cdot k+1}) - s) < tolS)$ ou $a = None$ **então**:
19 Tome $v = \frac{1}{n}\sum_m D_{m,k+1}$
20 **fim se**
21 **retorne** a e v
Fim

FONTE: A autora (2018)

O algoritmo é iniciado com o valor de impureza da amostra dada (linha 1), o que auxília que só seja escolhida uma divisão da amostra, se houver melhora no resultado atual. Os dois laços do algoritmo, linhas 4 e 5, efetuam a busca em todo o espaço possível de atributos e valores para encontrar a divisão da amostra de dados que possuir maior redução do erro.

Uma combinação atributo e valor, só é considerada para gerar os dois novos nós, se o tamanho de ambos os subconjuntos for pelo menos igual ao mínimo estabelecido (linha 8). Sendo considerado, é verificado se o conjunto minimiza a impureza, o que é realizado na linha 10.

A cada iteração, se os novos nós propostos melhorarem o estimador, o atributo e o seu valor de divisão são atualizados. Ao percorrer todas as divisões possíveis, é verificado se a melhoria é superior que a mínima estabelecida. Uma forma de verificar a variação da impureza é dada por:

$$\Delta R = R(Y) - (R(Y_e) + R(Y_d)),$$

que determina a diferença da impureza do nó atual e a soma da apurada em seus

nós filhos, que originam o ramo esquerdo e direito do nó pai, a impureza é dada pela equação 3.3.

Ao fim do algoritmo, retorna-se o atributo e o valor de divisão. No caso em que foi identificado um dos critérios para formar o nó terminal, o atributo retornado pelo algoritmo de seleção é $None$ e o valor é o estimado para a folha, dado pela equação 3.2.

Ainda existe a poda de uma árvore, que é uma forma de limitar o seu crescimento e reduzir a sua complexidade. O algoritmo de escolha já está preparado para esse processo, pelos parâmetros de $tolS$ e $tolN$. Por exemplo, ao se estabelecer o número mínimo de elementos do conjunto igual a um, não se aplica à poda, pois a árvore irá crescer até encontrar nós puros.

Os parâmetros assumidos no estudo são estabelecidos na Capítulo 4. De um modo geral, a matriz de dados (D) é composta pelas características conhecidas (x) de n tempestades e dos valores esperados para a característica que deseja-se prever (y).

Segundo Geurts, Ernst e Wehenkel (2006), o algoritmo padrão de árvores de decisão tem baixo custo computacional, o que o torna atraente para métodos de *ensemble*, em que existe a necessidade do desenvolvimento de vários preditores.

Dessa forma, a predição de uma variável regressora pode ser representada pelo método de aprendizado agrupado baseado em árvores de decisões pela FIGURA 12, em que são utilizadas m árvores e a função agupadora é a média das estimativas individuais.

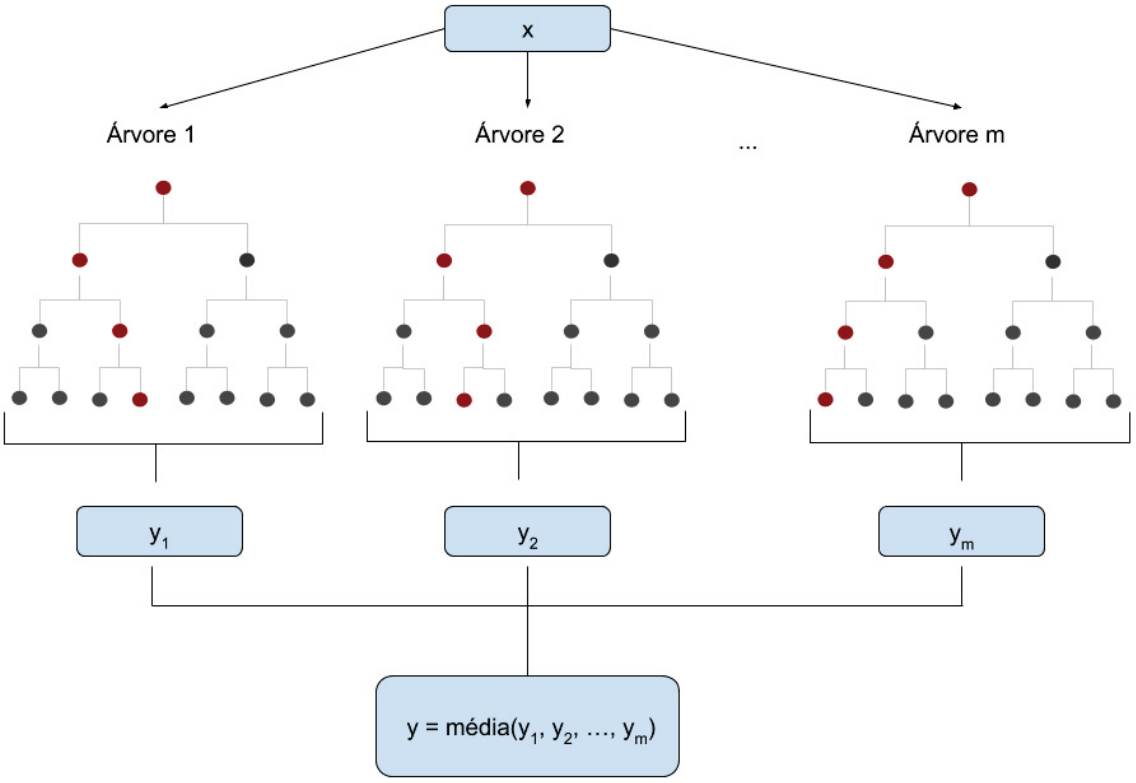
Nesse sentido, o processo para gerar cada uma dessas árvores varia de acordo com o método de conjunto. Os métodos utilizados nesse trabalho são apresentados nas seções 3.3.2 (Bagging), 3.3.3 (Random Forest) e 3.3.4 (Extra Trees).

3.3.2 Bagging

Bagging ou agregação *bootstrap* é um *ensemble* de reamostragem que foi desenvolvido por Breiman (1996). Nesse método, a partir de réplicas *bootstrap* do conjunto de dados, são gerados diversos preditores de um mesmo algoritmo, em que o preditor agregado é dado pela média dos individuais.

Nessa abordagem, a aleatoriedade ou variação dos preditores indivíduais é obtida, segundo Geurts, Ernst e Wehenkel (2006), pela construção de cada preditor por meio da amostragem *bootstrap* do conjunto original. Essa amostragem possui o mesmo tamanho do conjunto original, é uniforme e com reposição, o que implica que um mesmo exemplo pode aparecer mais de uma vez compondo o conjunto de

FIGURA 12 – PREDIÇÃO DE UM *ENSEMBLE* BASEADO EM ÁRVORE DE DECISÃO



FONTE: A autora (2018)

treinamento, enquanto outro pode não ser selecionado (OSHIRO, 2013).

Seja m o número de estimadores inviduais a serem treinados e D_p que representa a p-ésima amostra *bootstrap* gerada a partir do conjunto de dados (D), com $p = 1, .., m$, denota-se por estimador o algoritmo de aprendizagem utilizado e, desse modo, E_p representa a p-ésima função regressora gerada com amostra *bootstrap* correspondente e o estimador escolhido. Assim, o ALGORITMO 2 representa o processo de aprendizado do *Bagging*.

ALGORITMO 2 – BAGGING

Parâmetros: m, estimador e D
Início
1 **para** $p = 1$ até m **faça**:
2 Crie a amostra de bootstrap D_p
3 Treine o estimador E_p com a amostra D_p
4 **fim do para**
5 **retorne** o estimador agregado $\frac{1}{m}\sum_p E_p(x)$
Fim

FONTE: A autora (2018)

O método é "quase um procedimento dos sonhos para a computação paralela"(BREIMAN, 1996), uma vez que a geração de uma amostra não depende de outra, não necessitando de uma comunicação entre esses processos. Trata-se de "uma maneira relativamente fácil de melhorar um método existente"(BREIMAN, 1996), visto que tudo o que é necesssário é repetir a execução do mesmo algoritmo de amostragem (*bootstrap*) e um método de predição já existente, tantas vezes quantas forem desejadas, paralelamente ou não, e realizar a agregação das predições.

Segundo Brilhadori e Lauretto (2013), esse tipo de técnica permite que, com o aumento do número de preditores, seja possível reduzir a variância do conjunto, principalmente quando se trata de dados ruidosos. Ocorre para métodos instáveis, permitindo que tais métodos instáveis, mas bons, se aproximem mais da otimalidade. Entretanto, pode reduzir a precisão de procedimentos mais estáveis (BREIMAN, 1996).

Como apresentado anteriormente, árvores de decisão e rede neural são procedimentos do tipo instáveis. As redes neurais progridem mais lentamente quando comparadas às árvores de decisões que, de acordo com Koronacki, Ras e Wierzchon (2009), em comparação a outras técnicas de Aprendizado de Máquina, apresentam um rápido processamento na fase de treinamento do modelo.

Nesse contexto, o algoritmo Bagging refere-se à implementação a partir de árvores de decisões como preditores individuais, dado seu potencial de aprendizado e a rapidez nesse processo. Toma-se o estimador no ALGORITMO 2, como o CART. Utiliza-se, portanto, o crescimento de árvores padrão, com a seleção da divisão dos nós apresentada no ALGORITMO 1. A única alteração para cada estimador é a amostra inicial apresentada.

3.3.3 Random Forest

O *Random Forest* ou Floresta Aleatória proposto por Breiman (2001), "consiste em usar entradas selecionadas aleatoriamente ou combinações de entradas em cada nó para cultivar cada árvore". Segundo Geurts, Ernst e Wehenkel (2006), trata-se de um aprimoramento do *Bagging*, pois baseia-se na réplica *bootstrap* de amostragem. Entretanto, aplica-se restritamente ao agrupamento de árvores de decisões como preditores, visto que seu diferencial é justamente a inserção de aleatoriedade na divisão dos nós de cada árvore.

A aleatoriedade proposta nos preditores é introduzida avaliando-se apenas um subconjunto de atributos, em cada partição dos nós das árvores de decisão. Cada amostra avaliada na divisão é composta por k características, selecionadas sem substituição e distribuída uniformemente. Dos atributos dispostos, seleciona-se a melhor

ramificação (MOHAN; CHEN; WEINBERGER, 2011).

Dessa forma, o algoritmo *Random Forest* tem a mesma estrutura do *Bagging* e a seleção da melhor partição se mantém. Portanto, a diferença dos dois métodos está na amostra apresentada para a seleção da melhor ramificação. Enquanto o *Bagging* apresenta toda amostra para ser dividida, as Florestas Aleatórias selecionam um subconjunto aleatório de atributos para constituir o espaço de busca da divisão de cada nó.

Nessa seleção aleatória, cada característica tem a mesma probabilidade de ser escolhida e não pode ser selecionada mais de uma vez para o mesmo nó. Os atributos são escolhidos dentre todos os possíveis, para cada divisão, indiferente de ter sido escolhido ou não anteriormente. O aprendizado agrupado é obtido tomando-se a média dos regressores individuais.

A introdução da seleção aleatória de atributos no *ensemble* permite um ganho quanto a robustez em relação a ruídos (BREIMAN, 2001; LORENZETT; TELÖCKEN, 2016). Outra vantagem do método é que assim como o *Bagging* o método é "um algoritmo inerentemente paralelo"(MOHAN; CHEN; WEINBERGER, 2011), pois a construção de cada arvore é independente entre si.

Além dessas vantagens, Lorenzett e Telöcken (2016) afirmam que devido à característica "dividir para conquistar"do método, ele evita o sobreajuste ou *overffing* e, para diferentes conjunto de dados e possui boa taxa de acerto.

Nessa abordagem, com o aumento do número de árvores o erro de generalização converge a um limite, fazendo-se importante dois parâmetros, o número de árvores que comporão o modelo e a profundidade de cada árvore (BREIMAN, 2001).

3.3.4 Extra trees

Extra Trees ou *Extra Randomized Trees*, em português Árvores de decisões extremamente randomizadas ou aleatórias, foi proposta por Geurts, Ernst e Wehenkel (2006) e trata-se de um *ensemble* que introduz aleatoriedade tanto na escolha dos atributos como no ponto de corte para dividir um nó de árvore de decisão.

Quando comparado aos dois métodos já apresentados, distingue-se pelo uso de todos os exemplos do conjunto de dados, ao invés de réplicas *bootstrap*. Quanto à função de agrupamento dos preditores, também utiliza-se da média, como definido no ALGORITMO 3, onde m é o número de estimadores e D conjunto de dados.

Em termos de algoritmo, além da diferença quanto ao uso da amostra de dados, mostrado no ALGORITMO 3, as principais diferenças estão no crescimento das árvores.

Mantém o processo de seleção de atributos aleatórios incorporado pelo *Random Forest* na divisão de cada nó. Nesse processo, a divisão do nó é antecedida pela formação aleatória de um subconjunto de exemplos e atributos, ao invés de usar todo o conjunto de dados.

ALGORITMO 3 – EXTRA TREES

Parâmetros: m e D
Início
1 Tome D como a amostra de dados
2 **para** $p = 1$ até m **faça**:
3 Treine E_p pelo algoritmo CART aleatório com a amostra D
4 **fim do para**
5 **retorne** o estimador agregado $\frac{1}{m}\sum_p E_p(x)$
Fim

FONTE: A autora (2018)

Em adição, a técnica ainda incorpora a aleatoriedade na escolha dos valores para a divisão. O corte para um atributo numérico de entrada dado é aleatória e feito com base em um número de atributos selecionados, também, aleatoriamente, ou seja, o processo é realizado independente do rótulo desejado (GEURTS; ERNST; WEHENKEL, 2006). Essa alteração é ainda mais simples: seleciona-se aleatoriamente um atributo e um ponto de corte que satisfaça o número mínimo de dados em cada nó, ao invés de avaliar a melhor divisão para uma subamostra.

Em termos de viés e variância, Geurts, Ernst e Wehenkel (2006) defendem que a aleatorização no processo do ponto de corte pode permitir uma redução na variância, pois ela "pareceu ser responsável por uma parte significativa das taxas de erro dos métodos baseados em árvores". Enquanto isso, o uso do conjunto de dados original permite a minimização do viés.

Geurts, Ernst e Wehenkel (2006) afirmam ainda que "além da precisão, a principal força do algoritmo resultante é a eficiência computacional"e, assim, como o *Bagging* e o *Random Forest* permite a paralelização.

3.4 MÉTODOS BASEADOS EM MODELO LINEAR

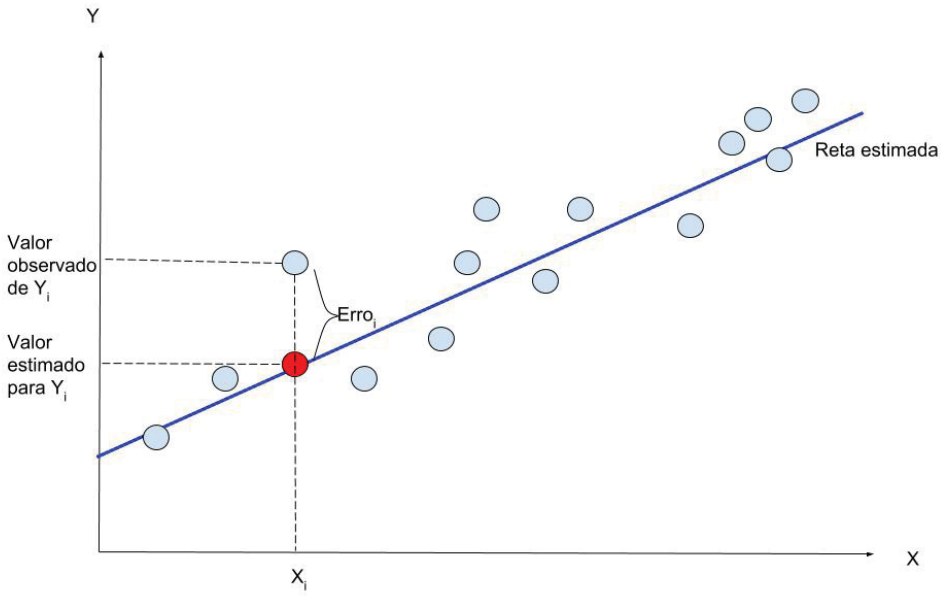
Os métodos de modelo linear podem se tratar de uma regressão linear simples ou múltipla. Diz-se que um modelo linear é simples quando baseia-se na relação entre duas váriaveis x e y, de modo que y é dependente de x. Nesse caso, deseja-se ajustar uma reta que melhor descreva a relação entre essas duas variáveis e permita estimar

o valor aproximado y dado por $\hat{y}$. Seja uma amostra com n exemplos, o valor estimado do exemplo i pode ser obtido pela seguinte equação:

$$\hat{y}_i = w_0 + w_1 x_i, \tag{3.5}$$

Nessa abordagem, objetiva-se encontrar valores para w_0 e w_1 que ajustem a melhor reta, produzindo o menor erro de predição de y com base nas observações x (WILKS, 2011), como mostra a FIGURA 13.

FIGURA 13 – MODELO LINEAR SIMPLES



FONTE: A autora (2018)

Existem diversos critérios de erros que podem ser minimizados e, neste trabalho, adota-se a soma dos erros quadrados, conhecido como método dos mínimos quadrados (MMQ). O erro consiste na diferença entre o valor esperado e o previsto. A partir da equação 3.5, obtem-se a expressão 3.6 a ser minimizada.

$$\sum_{i=1}^{n} [y - (w_0 + w_1 x_i)]^2 \tag{3.6}$$

Para o modelo de regressão linear o critério dos mínimos quadrados se reduz em determinar os valores de w_0 e w_1, que minimizam a expressão 3.6. Diferenciando 3.6 em relação a cada um dos coeficientes, linear e angular, respectivamente, e igualando a zero, tem-se as chamadas **equações normais** (RAWLINGS; PANTULA; DICKEY,

2001).

$$\sum_{i=1}^{n} y_i = (\sum_{i=1}^{n} 1)w_0 + (\sum_{i=1}^{n} x_1)w_1 \tag{3.7}$$

$$\sum_{i=1}^{n} x_i y_i = (\sum_{i=1}^{n} x_i)w_0 + (\sum_{i=1}^{n} x_1^2)w_1 \tag{3.8}$$

Assim, para minimizar a equação 3.6, é preciso resolver o sistema de equações formado por 3.7 e 3.8. O sistema pode ser escrito na forma matricial, dada abaixo.

$$\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} \tag{3.9}$$

Definindo-se uma matriz Y que contêm todos os rótulos para os n exemplos, outra w dos coeficientes e uma matriz $\phi(x)$, inversível, que contêm uma coluna unitária e uma outra com a variável regressora x.

$$Y = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}, w = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}, \phi(x) = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

O sistema de equações representado em 3.9 pode ser reescrito em relação as matrizes $\phi(x)$, w e Y, da sequinte forma:

$$w = (\phi(x)^T \phi(x))^{-1} \phi(x)^T Y \tag{3.10}$$

Desse modo, a resolução de um modelo linear simples usando o método dos mínimos quadrados, pode ser obtida pela a equação 3.10. Contudo, frequentemente, mais variáveis independentes são necessárias para estimar o valor de uma variável y, tornando o modelo simples insuficiente.

Segundo Wilks (2011), "as ideias para regressão linear simples generalizam-se facilmente para este caso mais complexo de regressão linear múltipla". Ao contrário do caso simples, possui mais de uma variável regressora, mas mantém-se uma única a ser estimada. Assim, toma-se k variáveis regressoras e, então, é possível reescrever a equação 3.5 de uma forma mais geral:

$$\hat{y}_i = w_0 + \sum_{j=1}^{k} w_j x_{ij} \tag{3.11}$$

O modelo simples pode ser reduzido a um caso especial do múltiplo, em que $k = 1$. Quanto aos $k + 1$ coeficientes w são denominados parâmetros de regressão, sendo que o w_0 pode ser chamado, ainda, de constante de regressão (WILKS, 2011).

Analogamente ao processo realizado para o modelo linear simples, as matrizes $\phi(x)$, w podem ser reescritas para o problema de $k + 1$ parâmetros a serem estimados.

$$\phi(x) = \begin{pmatrix} 1 & x_{11} & ... & x_{1k} \\ 1 & x_{21} & ... & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & ... & x_{nk} \end{pmatrix}, w = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{pmatrix} \tag{3.12}$$

As matrizes w e $\phi(x)$, reformuladas para um problema com k atributos, permitem a obtenção de uma formulação mais geral para o modelo linear e, mantém-se a equação 3.10 para determinar os valores dos parâmetros de regressão que minimizam a soma do quadrado dos erros.

O desenvolvimento apresentado para o MMQ e, assim, a equação 3.10 é bem semelhante a apresentada por Rawlings, Pantula e Dickey (2001). No entanto, pode ser encontrada em diversas referências, como Haykin (2007) e Huber e Ronchetti (2009), também utilizadas como base neste trabalho.

É importante salientar que os resíduos e são considerados com distribuição normal. No entanto, ao MMQ a normalidade não é necessária, pois mesmo em sua ausência, de acordo com Rawlings, Pantula e Dickey (2001), "as estimativas de mínimos quadrados são as melhores estimativas lineares não tendenciosas".

Os erros do modelo simples, assim como a regressão linear realizada, são facilmente visualizáveis, plotando-se uma reta e os pontos observados, como mostra a FIGURA 13. Entretanto, a dimensionalidade do problema cresce com o aumento do número de atributos para $k + 1$ dimensões. Com $k = 2$ ainda é possível representar a função regressora em uma superfície, mas para mais características, se perde essa facilidade.

Uma vantangem ao utilizar o MMQ é a tolerância obtida a pontos pouco distantes do valor observado. Contudo, como o método busca ajustar a função regressora à discrepâncias maiores, torna-se sensível a *outliers* (WILKS, 2011).

3.4.1 Theil Sen

A mediana das inclinações em pares de pontos foi proposta por Theil (1950) para regressão linear simples e, então, estendida para laços por Sen (1968), originando

o nome Theil Sen. Mais tarde, Dang et al. (2008) propuseram uma generalização do método para regressão linear múltipla, melhorando resultados já existentes na literatura.

Para o caso de modelo linear simples, o método baseia-se em determinar os coeficientes da reta, para todos os pares de exemplos (x_i, y_i) e (x_j, y_j), que satisfazem $x_i \neq x_j$ no intervalo $1 \leq j < i \leq n$ e são obtidos da combinação $C_{2,n}$ (SEN, 1968).

$$w_0 = med\left\{w_{0ij} = (y_j x_i - y_i x_j)/(x_i - x_j) : \forall (i,j) \in C_{2,n}\right\} \tag{3.13}$$

$$w_1 = med\left\{w_{1ij} = (y_i - y_j)/(x_i - x_j) : \forall (i,j) \in C_{2,n}\right\} \tag{3.14}$$

Assim, o coeficiente linear (3.13) e o angular (3.14) são obtidos calculando-se a mediana de cada um desses parâmetros, determinados para todos os pares de pontos da combinação $C_{2,n}$.

Para a generalização do modelo para um número k de atributos, é necessário um número m de pontos a serem combinados, de tal modo que permita determinar o valor dos $(k+1)$ coeficientes, ou seja, da matriz w.

Nesse sentido, a partir da escolha de m no intervalo fechado $(k+1, n)$ é possível determinar w pela equação 3.10, minimizando o critério de mínimos quadrados. Para isso, admite-se a matriz $\phi(x)_{Cm,n}$ inversível, que contêm as linhas da matriz $\phi(x)$ correspondentes ao subconjunto da amostra gerado para cada combinação $C_{m,n}$.

$$C_{m,n} = \begin{pmatrix} n \\ m \end{pmatrix} \tag{3.15}$$

O número de combinações obtidas em $C_{m,n}$ é denotada por N e, portanto, cada matriz $\phi(x)_{C_{m,n}}$ pode ser escrita como $\phi(x)_p$, com $p = 1, ..., N$. Dessa forma, uma maneira mais geral em substituição às equações 3.13 e 3.14 é dada por:

$$w = Mmed\left\{\hat{w}_p = (\phi(x)_p^T \phi(x)_p)^{-1} \phi(x)_p^T Y_p : \forall p = 1, ..., N\right\}, \tag{3.16}$$

onde $\hat{w}_p$ denota os pesos estimados para a combinação p. Nessa abordagem, como a dimensão de w é maior, a mediana univariada torna-se insuficiente e, portanto, adota-se a mediana multivariada ($Mmed$). A mediana espacial da amostra, baseia-se em maximizar a profundidade e, dessa forma, se a distribuição for simétrica representa o centro de simetria, como definida por Dang et al. (2008).

Assim, apresentados D como o conjunto de exemplos de toda a amostra de treinamento e m o número de exemplos utilizados na combinação $C_{m,n}$, o ALGORITMO 4 representa o processo do método Theil Sen.

ALGORITMO 4 – THEIL SEN

Parâmetros: N, $C_{m,n}$ e D
Início
1 **para** $p = 1$ até N **faça**:
2 Crie a subamostra D_p das combinações $C_{m,n}$
3 Calcule o $\hat{w}_p$ pelo MMQ de D_p
4 **fim do para**
5 $w = Mmed\{\hat{w}_p : p = 1, .., N\}$
6 **retorne** w
Fim

FONTE: A autora (2018)

No ALGORITMO 4, na linha 3 se dá a resolução do método dos mínimos quadrados para cada subconjuto da amostra (D_p), gerado pela combinação expressa na equação 3.15. Enquanto isso, na linha 5 determina-se os coeficientes pela mediana multivariada dos estimados com cada uma das subamostras. Por fim, a partir de w determinado, obtem-se $f(x)$ a função regressora como dado na equação 3.11, em que x representa qualquer variável regressora de k atributos.

Nessa abordagem, a escolha do parâmetro m como o número de coeficientes a serem determinados permite a maior robustez possível do algoritmo. Por outro lado, a escolha como o número de amostra permite a maior eficiência. Quando se tem m exatamente igual ao número de amostras, o modelo é equivalente ao MMQ (DANG et al., 2008). A escolha da variável m deve ser feita, com base na robustez e eficiência desejadas do algoritmo.

3.4.2 Bayesian Ridge

A abordagem bayesiana em modelos lineares de regressão consiste na incorporação do teorema de Bayes no processo de aprendizado, de modo que capture as suposições dos coeficientes do modelo (w), como uma distribuição prévia de probabilidade $p(w)$ e o efeito dos dados observados (D), dado pela probabilidade condicional $p(D \mid w)$. A regra de Bayes pode ser escrita como dada na equação 3.17 (BISHOP, 2006):

$$p(w \mid D) = \frac{p(D \mid w)p(w)}{p(D)} \tag{3.17}$$

De acordo com Neal (2012), a função de verossimilhança, dada por $p(D \mid w)$, permite capturar o impacto das observações dos dados, como função de w, ou seja, mensura a probabilidade dos dados para a diversidade de configurações do vetor de

parâmetros do modelo. Com isso, é possível reescrever a regra de Bayes em palavras, da seguinte forma:

$$\text{posteriori} \propto \text{verossimilhança} \times \text{priori}$$

O denominador da equação 3.17 representa uma constante de normalização e não depende de w, e "é comumente ignorada, uma vez que é irrelevante para o primeiro nível de inferência"(MACKAY, 1992). Assim, a representação do teorema de Bayes é dada em relação a priori e a verossimilhança, em que a priori é dada pela suposição prévia dos parâmetros.

Nesse contexto, é de interesse que os parâmetros sejam ajustados a maximizar a probabilidade do conjunto de dados, que pode ser obtido por um estimador de máxima verossimilhança.

Uma abordagem bastante utilizada em aprendizado de máquina, segundo Bishop (2006), é o uso do logaritmo negativo da verossimilhança, como função de erro. Dessa forma, maximizar a probabilidade é equivalente a minimizar o erro do modelo.

A construção de um modelo linear objetiva prever valores alvos da variável dependente, dadas observações das variáveis independentes, com base nos dados de treinamento. Com esse fim, sejam x o dado observado e o valor alvo y, a estimativa pode ser realizada a partir de uma função dos parâmetros e as observações ($f(x, w)$), como apresentado na equação 3.11, equivalente a $\phi(x)w$, com w e $\phi(x)$ dados na expressão 3.12.

Na abordagem bayesiana a aprendizagem se dá expressando as incertezas, a partir do uso da probabilidade. A incerteza quanto ao valor esperado pode ser obtida assumindo uma distribuição de probabilidade. Com esse propósito, assumindo-se que y tem distribuição gaussina, em que sua média é expressa por $f(x, w)$, toma-se a variância inversa, como parâmetro de precisão, β. Assim, a função de verossimilhança é definida pela equação 3.18, com w o vetor de $k + 1$ elementos.

$$P(y \mid X, w, \beta) = \mathcal{N}(y \mid f(x, w), \beta^{-1}) \tag{3.18}$$

Quanto a priori, considerando a distribuição gaussiana, seja α a precisão da distribuição prévia de w, tem-se:

$$P(w \mid \alpha) = \mathcal{N}(w \mid 0, \alpha^{-1}I). \tag{3.19}$$

Portanto, a distribuição posterior é dada pelo teorema de Bayes, como o produto da prévia e da função de verossimilhança. Devido à escolha de uma distribuição prévia

Gaussiana conjugada, a posterior também é Gaussiana.

$$p(w \mid X, y, \alpha, \beta) \propto p(y \mid X, w, \beta)p(w \mid \alpha) \tag{3.20}$$

Tomando β como uma constante conhecida, tem-se

$$p(w \mid y) \propto p(y \mid w)p(w), \tag{3.21}$$

em que a função de verossimilhança, então dada por $p(y \mid w)$, é a exponencial de uma função quadrática de w.

Com isso, a distribuição posterior pode ser obtida completando quadrado na exponencial e, a partir do resultado padrão de uma gaussina normalizada obtem-se o coeficiente de normalização, como demonstrado por Bishop (2006).

Desse modo, é possível obter a distribuição posterior diretamente da equação 3.22, em que S e m_N são definidos nas equações 3.23 e 3.24.

$$p(w \mid y) = \mathcal{N}(w \mid m_N, S^{-1}) \tag{3.22}$$

$$S = \alpha I + \beta X^T X \tag{3.23}$$

$$m_N = \beta S^{-1} X^T y \tag{3.24}$$

O vetor de pesos, posterior máximo, é dado pela média da distribuição posterior, pois devido a ser gaussiana a média (m_N) e o modo coincidem. A média preditiva é representada pela combinação ponderada das variáveis alvo, de modo que é possível escrever $f(x, m_N)$ como apresentado abaixo.

$$f(x, m_N) = m_N^T \phi(x) = \beta \phi(x)^T S_N^{-1} X^T y = \sum_{i=1}^{n} \beta \phi(x)^T S_N^{-1} \phi(x_i) y_i. \tag{3.25}$$

Para uma abordagem do modelo linear totalmente bayesiano, existe a inclusão da distribuição prévia de α e β, que são chamados de hiperparâmetros, pois se aplicam no controle da distribuição dos parâmetros. Dessa forma, a marginalização de w, α e β define a distribuição preditiva, como dada abaixo.

$$p(\hat{y} \mid x, D) = \int\int\int p(\hat{y} \mid x, w, \beta)p(w \mid D, \alpha, \beta)p(\alpha, \beta \mid D)\mathrm{d}w\mathrm{d}\alpha\mathrm{d}\beta \tag{3.26}$$

A integração analítica da expressão, visando a marginalização completa em relação a todas as variáveis é inviável. Para contornar esse problema, adota-se o procedimento de aproximação de evidências, também conhecido como Bayes empírico, máxima verossimilhança generalizada ou de tipo 2. A partir da função de verossimilhança marginal maximizada em relação a integração de w, são definidos valores específicos aos hiperparâmetros. Da regra de Bayes, tem-se:

$$p(\alpha, \beta \mid y) \propto p(y \mid \alpha, \beta)p(\alpha, \beta). \tag{3.27}$$

Seja $p(\alpha, \beta)$ relativamente monótona, os valores de α e β podem ser estimados maximizando a função de verossimilhança marginal, pela integração em relação a w, como apresenta a equação 3.28.

$$p(y \mid \alpha, \beta) = \int p(y \mid w, \beta)p(w \mid \alpha)\mathrm{d}w \tag{3.28}$$

Definida a distribuição prévia pela equação 3.19 e $p(y \mid w, \beta)$ com distribuição gaussiana, como dado na equação 3.29, com a expressão relativa ao erro quadrático dada por 3.30.

$$P(y \mid w, \beta) = \sum_{i=1}^{n} ln\mathcal{N}(y \mid w^T\phi(x_i), \beta^{-1}) = \frac{n}{2}ln\beta - \frac{n}{2}len(2\pi) - \beta E_D(w) \tag{3.29}$$

$$E_D(w) = \frac{1}{2}\sum_{i=1}^{n}(y_i - w^T\phi(x_i))^2 \tag{3.30}$$

A integral pode ser abordada com o mesmo procedimento descrito para a obtenção da distribuição preditiva dada na equação 3.28. Obtém-se a função de evidência como dada na equação 3.31, com $E(w)$ dado em 3.32.

$$p(y \mid \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}}\left(\frac{\alpha}{2\pi}\right)^{\frac{k}{2}}\int exp(-E(w))dw \tag{3.31}$$

$$E(w) = \beta E_D(w) + \alpha E_w(w) = \frac{\beta}{2}\left\|y - \phi(x)w\right\|^2 + \frac{\alpha}{2}w^Tw \tag{3.32}$$

Com isso, pode-se estabelecer que $E(w)$ assume a forma da função da soma do erro quadrático regularizada, apresentada por Bishop (2006), em que α e β definem o parâmetro de regularização.

Dessa forma, a média preditiva como definida na equação 3.24, corresponde a matriz da segunda derivada de $E(w)$, ou seja a matriz hessiana, onde S é dado

pela equação 3.23. Com base na função de evidência, é possível obter o logaritmo da verossimilhança marginal dado pela equação 3.33.

$$ln[p(y \mid \alpha, \beta)] = \frac{k}{2} ln\alpha + \frac{n}{2} ln\beta - E(m_N) + \frac{1}{2} ln \mid S^{-1} \mid - \frac{n}{2} ln(2\pi), \tag{3.33}$$

$$E(m_N) = \frac{\beta}{2} \left\| y - \phi(x) m_N \right\|^2 + \frac{\alpha}{2} m_N^T m_N.$$

Com o objetivo de determinar os hiperparâmetros que implicam na maximização da função de log-verossimilhança, a equação do autovetor é definida pela equação 3.34, de forma que a matriz hessiana possui autovalores $\alpha + \lambda_i$.

$$(\beta X^T X) u_i = \lambda_i u_i \tag{3.34}$$

Com isso, o termo relativo a S da equação 3.33 pode ser derivado em relação a α e β, conforme as equações 3.35 e 3.36, com $\frac{\mathrm{d}}{\mathrm{d}\beta} \lambda_i = \frac{\lambda_i}{\beta}$.

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} ln \mid S \mid = \frac{\mathrm{d}}{\mathrm{d}\alpha} ln \prod_i (\alpha + \lambda_i) = \sum_i \frac{1}{\alpha + \lambda_i} \tag{3.35}$$

$$\frac{\mathrm{d}}{\mathrm{d}\beta} ln \mid S \mid = \frac{1}{\beta} \sum_i \frac{\lambda_i}{\alpha + \lambda_i} \tag{3.36}$$

Os pontos estacionários em relação a α e β, respectivamente, satisfazem as equações 3.37 e 3.38:

$$0 = \frac{k}{2\alpha} - \frac{1}{2} m_N^T m_N - \frac{1}{2} \sum_i \frac{1}{\lambda_i + \alpha}, \tag{3.37}$$

$$0 = \frac{n}{2\beta} - \frac{1}{2} \left\| y - \phi(x) m_N \right\|^2 - \frac{1}{2\beta} \sum_i \frac{\lambda_i}{\alpha + \lambda_i}. \tag{3.38}$$

A quantidade γ é escrita conforme a equação 3.39 e os hiperparâmetros pelas equações 3.40 e 3.41.

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}, \tag{3.39}$$

$$\alpha = \frac{\gamma}{m_N^T m_N} \tag{3.40}$$

$$\beta = \frac{n - \gamma}{\left\| y - \phi(x) m_N \right\|^2} \tag{3.41}$$

Com essas definições é possível apresentar a estrutura de um modelo linear bayesiano, que é dado de forma iterativa, como pode ser observado no ALGORITMO 5.

ALGORITMO 5 – BAYESIAN RIDGE

Parâmetros: $niter$, tol, X e y
Início
Inicialize $\beta = \frac{1}{Var(y)}$, $\alpha = 1$, $w = \beta(\alpha I + \beta X^T X)^{-1} X^T y + tol$.
Decompõe X e obtenha os autovalores (λ)
para $iter = 1$ até $niter$ **faça**:
    Calcule m_N
    Calcule γ
    Determine $\hat{\beta}$ e $\hat{\alpha}$
    se $w - m_N < tol$ **faça**:
        pare e **retorne** w
    senão:
        Atualize $w = m_N$, $\beta = \hat{\beta}$ e $\alpha = \hat{\alpha}$
    fim do se
fim do para
retorne w
Fim

FONTE: A autora (2018)

Inicialmente, são tomados valores para α e β, determinando-se os autovalores em relação a matriz de dados observados.

A partir dessas informações, dado um número de iterações ($niter$), o procedimento iterativo é iniciado, em cada passo é calculada a média preditiva (eq. 3.24), γ (eq. 3.39) e estima-se o valor dos hiperparâmetros como apresentado nas equações 3.40 e 3.41.

As equações, tanto para a estimativa do parâmetro w, bem como de α e β são definidas de acordo com a estimativa de máxima verossimilhança, de modo que é equivalente a redução do erro médio quadrático em cada iteração.

Nesse sentido, é preciso estabelecer até que momento é vantajoso executar o processo iterativo. Além do número máximo de iterações assumido, ainda toma-se uma tolerância (tol), que permite que o processo continue apenas quando há uma variação significativa nos resultados anteriores. Ao fim do algoritmo obtem-se o vetor de parâmetros do modelo linear, que permite a previsão de novas entradas de dados pela equação 3.11.

Nessa abordagem, a inserção da probabilidade permite expressar as incertezas do modelo, dando um novo significado ao teorema de Bayes.

Com isso, os métodos aplicados na pesquisa são apresentados em dois grupos. Ambos para a abordagem do problema de regressão, pois a estimativa desejada

classifica-se como a previsão de um valor numérico. Os algoritmos selecionados de aprendizado agrupado são o *Bagging*, *Random Forest* e *Extra Trees*, métodos de aprendizado agrupado todos baseados no algoritmo de árvore de decisão. Já os modelos lineares são o *Bayesian Ridge*, baseado no teorema de Bayes e o *Theil Sen* que aplica a mediana multivariada em sua abordagem, ambos resistentes a dados ruidosos, ou seja, robustos.

Todos os processos que permeiam a construção dos estimadores avaliados na pesquisa, a partir dos cinco métodos dissertados, são apresentados no Capítulo 4.

4 MATERIAIS E MÉTODOS

O problema de pesquisa se constitui da previsão de um valor numérico e, dessa forma, classifica-se em regressão. Os algorimos selecionados para desenvolvimento desse trabalho são apresentados no seção 3.2 e pertencem a dois grupos, o aprendizado agrupado e modelos lineares robustos, técnicas de rápido processamento e resistentes a dados ruidosos.

As técnicas de AM se aplicam em diferentes áreas e problemas o que qualifica a técnica para um problema específico são os processos que permeiam a construção do modelo de aprendizado. Esses são definidos pelos dados utilizados no treinamento e sua configuração, como nos parâmetros de cada algoritmo.

4.1 MATERIAIS

O sul e sudeste do Brasil, são regiões frequentemente afetadas por eventos severos, com precipitação acima da média nacional (CALHEIROS, 2008). Nesse contexto, a aplicação da pesquisa se dá por meio de dados obtidos para essa região.

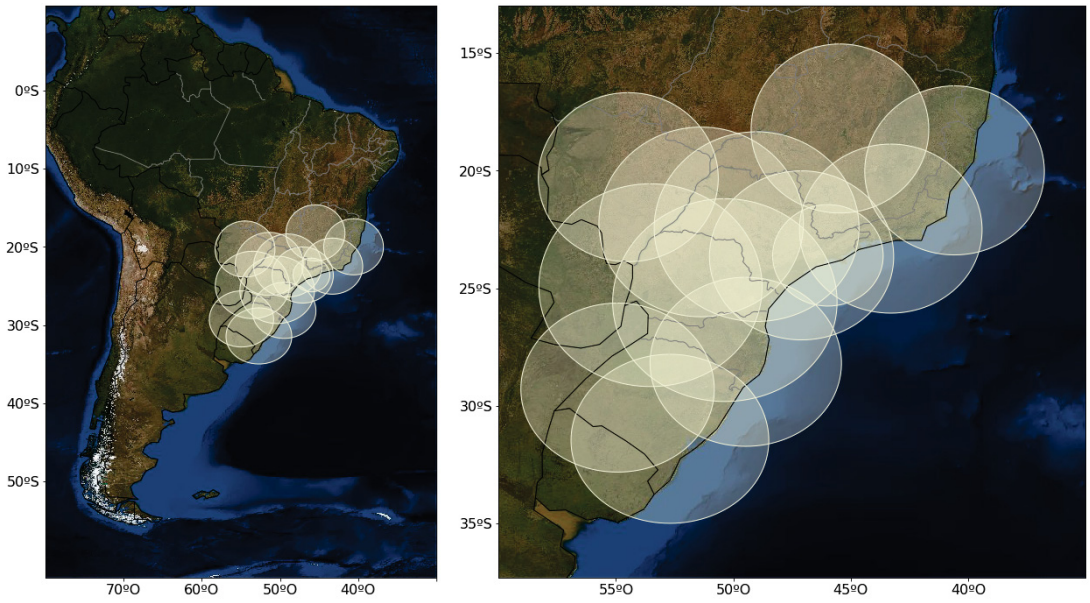
Os principais dados utilizados neste trabalho são provenientes da execução do software TITAN, com um mosaico de dados de radares meteorológicos. A área de estudo e a abrangência de cada radar que compõe o mosaico são apresentadas na FIGURA 14.

Os dados fornecidos para o treinamento e teste das técnicas de AM correspondem ao histórico de células de tempestades identificadas e acompanhadas pelo TITAN, no período de 15 de agosto de 2016 à 9 de agosto de 2017.

O histórico de dados obtidos a partir do TITAN, possui resolução temporal de 10 minutos, como descrito no subseção 2.4.3, onde para cada célula foram extraídas do software as características apresentadas no QUADRO 3.

Para as tempestades que possuem pelo menos 20 minutos de histórico, desde a sua primeira identificação, ainda tem-se as previsões do TITAN quanto a localização do centróide e o tamanho dos eixos da elipse de tempestade para os 60 minutos subsequentes, pois somenete a partir deste histórico mínimo a previsão é realizada pelo software.

FIGURA 14 – LOCALIZAÇÃO DOS RADARES PRESENTES NO MOSAICO DE DADOS DO ESTUDO



(a) Radares na região Sul/Sudeste do Brasil

(b) Aproximação da região da estudo

FONTE: A autora (2018)

4.1.1 Especificações dos conjuntos de dados

A separação dos dados é realizada levando-se em consideração o tempo que é possível estimar o deslocamento na mesma resolução temporal que o TITAN e, ainda, ter dados identificados para realizar a avaliação dos resultados, limitando-se a tempestades com pelo menos 20 minutos de histórico.

Seja n o número de históricos de tempestades no conjunto de dados, em que t representa o tempo, em minutos, de cada histórico, o período máximo que cada tempestade pode ser usada para previsão é dado por:

$$pmax_i = t_i - 20, \forall i = 1, ..., n \tag{4.1}$$

Assim, pode-se assumir um histórico mínimo necessário para cada tempo de previsão, conforme mostra a TABELA 1.

Desse modo, toma-se o conjunto de históricos de todas as tempestades, com mais de vinte minutos de duração, denotado por D, com n tempestades distintas. Os conjuntos de dados para cada período são denotados por $D_{(p)}$, em que p representa o tempo de previsão, ou seja, $p = 10, 20, ..., 60$.

Como foi discutido no Capítulo 3, para técnicas de aprendizado supervisionado

TABELA 1 – TEMPO MÍNIMO DE HISTÓRICO PARA CADA PREVISÃO

Tempo de previsão	Tempo minímo de histórico
10 minutos	30 minutos
20 minutos	40 minutos
30 minutos	50 minutos
40 minutos	60 minutos
50 minutos	70 minutos
60 minutos	80 minutos

FONTE: A autora (2018)

é necessária a divisão do conjunto de dados em treinamento e teste. O processo utilizado para compor esses conjuntos para cada p é dado pelo ALGORITMO 6. Assumindo a proporção 80% e 20% para os conjuntos de aprendizado, denotam-se $D_{T(p)}$ o conjunto de treinamento e $D_{t(p)}$ de teste, para o período p de até uma hora.

Para garantir que uma mesma tempestade pertence ao mesmo conjunto, teste ou treinamento, para todos os períodos t, toma-se a união dos conjuntos, partindo do período mais distante ao mais próximo, temporalmente, do período inicial de previsão T.

ALGORITMO 6 – PARTIÇÃO DOS CONJUNTOS DE DADOS

Parâmetro: D
Início
para $p = 60$ até 10 ao passo de -10 **faça**:
  se $p = 60$ **então**:
    $D_{(p)} = \{D_i \mid pmax_i \geq p, \forall i = 1, ..., n\}$
    atribua a $D_{T(p)}$ 80% dos dados aleatoriamente
    atribua a $D_{t(p)}$ os 20% restante dos dados
  senão:
    $D_{(p)} = \{D_i \mid pmax_i = p, \forall i = 1, ..., n\}$
    atribua a $D_{T(p)}$ 80% dos dados aleatoriamente
    atribua a $D_{t(p)}$ os 20% restante dos dados
    $D_{T(p)} = D_{T(p)} \cup D_{T(p-10)}$
    $D_{t(p)} = D_{t(p)} \cup D_{t(p-10)}$
fim do para
retorne os conjuntos $D_{T(p)}$ e $D_{t(p)}, \forall p$
Fim

FONTE: A autora (2018)

Considerando a divisão desenvolvida no ALGORITMO 6, as quantidades de células são dadas na TABELA 2. O tempo de ínicio de cada previsão é denotado por T, que é antecedido por pelo menos 20 minutos de dados históricos e utilizado para gerar até uma hora de previsão.

Para históricos maiores que 80 minutos, T é selecionado aleatoriamente, de modo que mantenha pelo menos 60 minutos subsequentes de dados identificados.

TABELA 2 – QUANTIDADE DE CÉLULAS DE TEMPESTADES POR CONJUNTO DE APRENDIZADO

Tempo de previsão	Treinamento	Teste
T + 10 minutos	40407	10109
T + 20 minutos	24789	6200
T + 30 minutos	14774	3696
T + 40 minutos	8881	2222
T + 50 minutos	5635	1410
T + 60 minutos	3592	899

FONTE: A autora (2018)

Isso, para não perder a viabilidade de avaliação dos modelos. Além das informações de radar processadas pelo TITAN, o processo de aprendizado também utilizou dados de descargas elétricas atmosféricas, cuja seleção é apresentada na subseção 4.1.2.

4.1.2 Seleção dos dados de descargas elétricas atmosféricas

Os dados de descargas elétricas atmosféricas fornecidos para a pesquisa são provenientes da Rede Integrada Nacional de Detecção de Descargas Atmosféricas (RINDAT), que é mantida pelo SIMEPAR em cooperação com Furnas, Companhia Energética de Minas Gerais (CEMIG) e Instituto Nacional de Pesquisas Atmosféricas (INPE).

As descargas são detectadas pelos sensores e, então, processadas. As informações são geradas e armazenadas. Tratam-se desde a localização geográfica e temporal das ocorrências, como também características como a polaridade e o pico de corrente do raio (RINDAT, 2018).

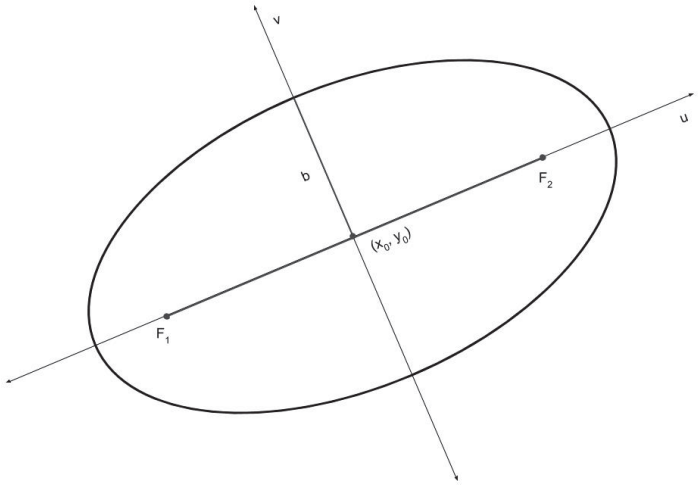
Para fazer uso desses dados juntamente com os do TITAN, se faz necessário identificar os dados de descargas elétricas que pertencem a cada célula, agregando os atributos relacionados a eletricidade da tempestade.

Para isso, a tempestade é considerada como uma elipse, como apresenta a FIGURA 15, em que o TITAN fornece as informações da posição do centróide da elipse, e o tamanho do eixo maior e menor, a partir desses dados é possível desenvolver a previsão das mesmas caracteríscas com as técnicas propostas e gerar as elipses previstas.

Tomando-se a o tamanho do semieixo maior e b do menor, F_1 e F_2, os focos da elipse, pelas definições de elipse, obtem-se:

$$| \vec{F_1} | = | \vec{F_2} | = \sqrt{a^2 - b^2} \tag{4.2}$$

FIGURA 15 – ELIPSE



FONTE: A autora (2018)

Sejam $\vec{F_1}$ e $\vec{F_2}$ a distância dos focos ao centróide da elipse, pela simetria das elipses, todo ponto pertencente a elipse, satisfaz a equação (SAFIER, 2009):

$$| \overrightarrow{PF_1} | + | \overrightarrow{PF_2} | = 2a, \tag{4.3}$$

onde $| \overrightarrow{PF_1} |$ e $| \overrightarrow{PF_2} |$ representam as distâncias do ponto P, a cada um dos focos. Com base na equação 4.3, qualquer ponto pertencente à elipse ou contido em sua região interna, deve satisfazer a equação 4.4.

$$| \overrightarrow{PF_1} | + | \overrightarrow{PF_2} | \leq 2a \tag{4.4}$$

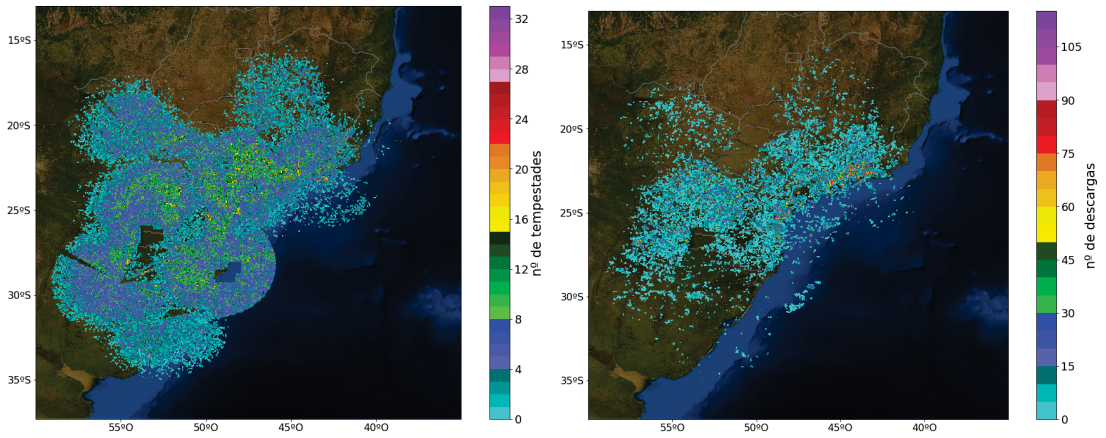
Dessa forma, para cada célula de tempestade C, é extraído um conjunto de dados de descargas elétricas atmosféricas (R), de modo que R é constituido pelo histórico de descargas que estão na região retangular formada pelos valores máximo e mínimo de x e y da elipse, em coordenadas cartesianas e, ainda, que ocorreram num espaço temporal dos 10 minutos que antecedem a tempestade.

Os focos de cada elipse C são determinados, e os pontos de descargas do conjunto R, que satisfazem a equação 4.4, são selecionados para comporem a informação da célula de tempestade.

Nesse sentido, a distribuição geográfica das células de tempestades e das descargas elétricas que compõem o conjunto de dados, podem ser visualizadas na FIGURA 16.

As informações relativas aos raios são incorporadas ao conjunto na forma de três atributos. O primeiro trata do número de descargas positivas identificadas para a

FIGURA 16 – DISTRIBUIÇÃO GEOGRÁFICA DOS DADOS



(a) Células de tempestades (b) Descargas elétricas atmosféricas

FONTE: A autora (2018)

célula de tempestade, o segundo do número de negativas e o terceiro o número total.

4.1.3 Conjunto de validação

A partir das informações apresentadas, estabelecidos os critérios relativos a formação dos conjuntos de dados, a agregação de informação de raio e a seleção do tempo inicial de previsão (T), é possível observar que a partição dos dados para cada processo de aprendizado, apresentada no ALGORITMO 6, é feita independente da data em que as células foram identificadas. Isso implica que podem existir células muito próximas temporal e espacialmente, em conjuntos distintos.

Levando-se em consideração que tanto a região de estudo, quanto o período são amplos, espera-se que isso não interfira no aprendizado, ou na capacidade de generalização.

Para verificar se de fato a disposição dos dados dessa forma não causa perda na generalização dos modelos, optou-se por tomar mais um conjunto de dados, denominado conjunto de validação.

Esse conjunto é gerado pelos mesmos critérios apresentados ao conjunto geral, mas sem a necessidade da divisão em treinamento e teste. Os dados são do período de 10 de agosto de 2017 à 31 de janeiro de 2018, e as quantidades de células para cada período de previsão são mostradas na TABELA 3.

Com base nas Tabelas 2 e 3, a proporção do novo conjunto de dados em relação ao geral pode ser visualizada na FIGURA 17. Pode-se verificar que o conjunto de validação representa uma proporção de metade do conjunto geral utilizado

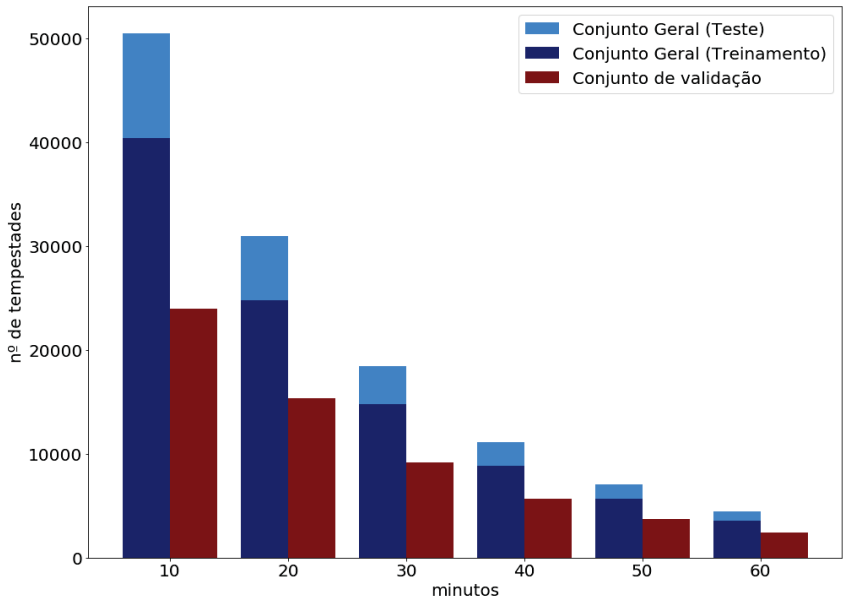
TABELA 3 – QUANTIDADE DE CÉLULAS DE TEMPESTADES POR CONJUNTO

Tempo de previsão	Número de células
T + 10 minutos	24005
T + 20 minutos	15336
T + 30 minutos	9132
T + 40 minutos	5667
T + 50 minutos	3689
T + 60 minutos	2389

FONTE: A autora (2018)

no aprendizado. Tal número é esperado, já que o primeiro conjunto é composto por aproximadamente 1 ano de dados, enquanto o segundo, aproximadamente 6 meses. Portanto, o conjunto selecionado representa uma proporção significativamente grande em relação ao número de dados utilizados para o treinamento das técnicas.

FIGURA 17 – NÚMERO DE TEMPESTADES POR CONJUNTO E PERÍODO DE PREVISÃO



FONTE: A autora (2018)

Por fim, o objetivo dessa nova amostra é de validar os modelos, testanto-os com dados totalmente externos aos processos de treinamento e teste já avaliados.

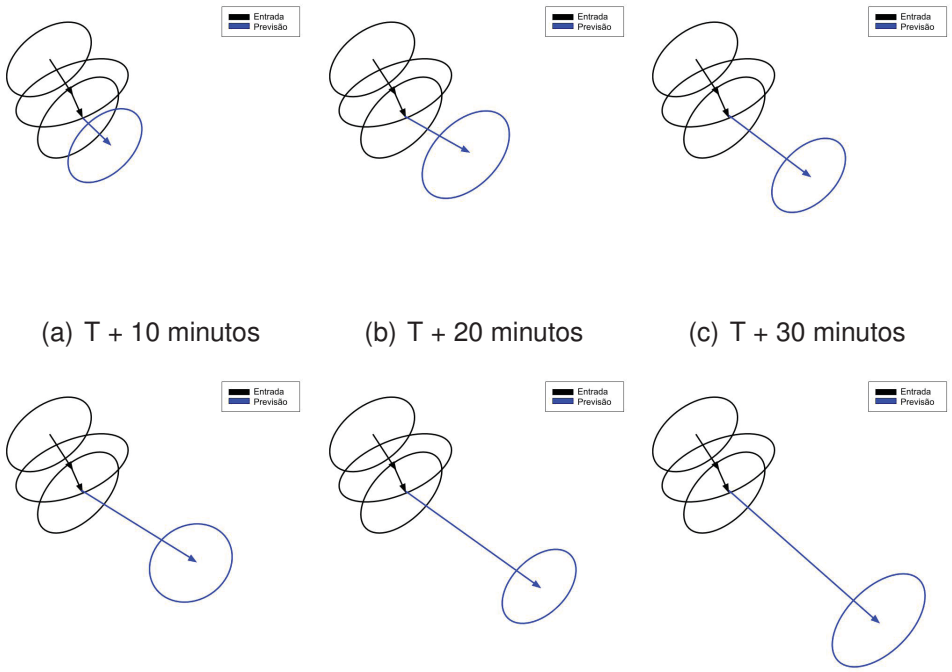
4.2 MÉTODOS

Os metódos que permeiam o desenvolvimento de uma pesquisa devem ser escolhidos de acordo com os dados e objetivos, visando a viabilidade do processo. Como o principal objetivo deste trabalho está relacionado ao acompanhamento de

tempestades em um horizonte de curtíssimo prazo, o deslocamento de tempestades é verificado e, então, é calculada a expansão dos modelos para abranger a estimativa dos tamanhos dos eixos da elipse de tempestade.

Por meio da previsão do deslocamento vertical e horizontal da célula e o tamanho dos eixos maior e menor, para cada um dos períodos, é possível gerar uma rota prevista da tragetória da tempestade. Para isso, toma-se o centróide previsto, com base na posição em T e na previsão do deslocamento em relação ao norte e em relação ao leste. A FIGURA 18 representa a elipse prevista em cada período de tempo, composta pelas informações estimadas da posição do centróide e dos tamanhos dos eixos.

FIGURA 18 – EXEMPLO DA PREVISÃO DO DESLOCAMENTO PARA CADA PERÍODO



(a) T + 10 minutos (b) T + 20 minutos (c) T + 30 minutos

(d) T + 40 minutos (e) T + 50 minutos (f) T + 60 minutos

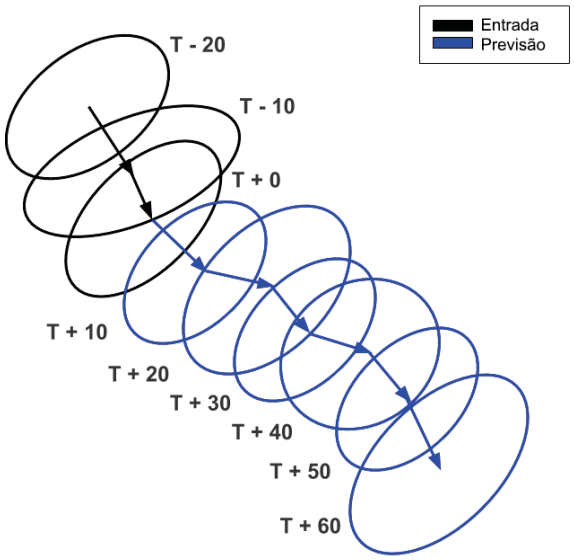
FONTE: A autora (2018)

A rota prevista é obtida a partir de cada uma das previsões para os períodos individuais, como mostra a FIGURA 19.

É possível verificar que a previsão proposta, assim como o TITAN, mantém a informação de orientação da tempestade constante, prevendo apenas as outras informações relativas à elipse. Para a obtenção dos modelos que gerem essas estimativas, existem dois importantes passos, listados a seguir.

1. Selecionar as características das tempestades para o processamento das técnicas; e
2. Parametrizar cada uma das técnicas propostas para o problema.

FIGURA 19 – EXEMPLO DA ROTA DE DESLOCAMENTO PREVISTA



FONTE: A autora (2018)

A linguagem computacional Python foi adotada nessa pesquisa pois se apresenta como um ambiente propício à preparação de dados e desenvolvimento de metodologias de AM. Devido a sua popularidade, proporciona diversas ferramentas, pois possui um amplo desenvolvimento de bibliotecas com código aberto e documentação (HARRINGTON, 2012). As principais ferramentas utilizadas são apresentadas a seguir:

- **Pandas:** Módulo para manipulação de tabelas de dados. Conta com várias funções gráficas e estatísticas para análise de dados (MCKINNEY, 2011). Utilizada para toda a manipulação dos dados e análise de resultados;
- **Scikt-Learn:** Módulo de AM que ”integra uma ampla gama de algoritmos de aprendizado de máquina de ponta, para problemas supervisionados e não supervisionados” (PEDREGOSA et al., 2011), com rotinas de pré-processamento à pós-processamento. Aplicada para o processamento das técnicas e manipulação dos algoritmos; e
- **Pyart (*Python ARM Radar Toolkit*):** Módulo de leitura, visualização e análise de dados de radares meteorológicos (HELMUS; COLLIS, 2016). Abordada na visualização e controle de qualidade dos dados de radares utilizados pelo TITAN.

Definido o ambiente de desenvolvimento dos métodos, ainda é preciso estabelecer como avaliar se um modelo é bom ou não. Com esse objetivo, adotam-se algumas métricas de avaliação das estimativas, apresentadas na subseção 4.2.1.

Dadas as ferramentas para mensuração do desempenho de um modelo, a busca por um bom conjunto de atributos e parâmetros, pode ser realizada conforme apresentado nas subseções 4.2.2 e 4.2.3, respectivamente.

4.2.1 Métricas para avaliação de desempenho

A avaliação da acurácia dos modelos é importante e permite não apenas avaliar o desempenho de um modelo para um determinado problema mas, também, comparar diferentes métodos para um mesmo problema.

Nesse sentido, com base em Rawlings, Pantula e Dickey (2001) e Wilks (2011), algumas métricas de erro são definidas. Para isso, é necessário estabelecer o que denomina-se por erro. Nesta pesquisa, o erro de previsão ou resíduo (ϵ), é definido como a diferença entre o valor observado (y) e o valor estimado ($\hat{y}$), como mostra a equação 4.5.

$$\epsilon(y, \hat{y}) = y - \hat{y} \tag{4.5}$$

Considerando -se i, que representa o i-ésimo exemplo do conjunto de amostra, formado por n exemplos, o erro médio absoluto (EMA) pode ser estimado pela equação 4.6.

$$EMA = \frac{1}{n} \sum_{i=1}^{n} \mid \epsilon(y_i, \hat{y}_i) \mid \tag{4.6}$$

O erro médio absoluto é uma métrica bem conhecida e comum na literatura. Permite mensurar o erro médio no conjunto, independente se a estimativa ocorreu acima ou abaixo do valor esperado, é relevante apenas a variação entre esses valores.

A soma do erro quadrático (SEQ) é utilizado no Método dos Mínimos Quadrados, apresentado na seção 3.4 e é dado pela equação 4.7. A partir dela, é possível definir a raiz do erro médio quadrático (REMQ), dado pela equação 4.8.

$$SEQ = \sum_{i=1}^{n} \epsilon(y_i, \hat{y}_i)^2 \tag{4.7}$$

$$REMQ = \sqrt{\frac{1}{n} SEQ} \tag{4.8}$$

O quadrado da diferença permite que os erros maiores ganhem maior peso. Estimando-se o EMA e o REMQ, pela diferença entre esses erros é possível expressar

a variação dos erros, em relação ao erro médio absoluto. Quanto menores forem os valores obtidos por essas métricas, como também a variação entre elas, mais confiável é a estimativa obtida.

Outra métrica bastante utilizada em problemas de regressão é o coeficiente de determinação (R^2), representado pela equação 4.9:

$$R^2 = 1 - \frac{SEQ}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{4.9}$$

onde $\overline{y}$ trata-se do valor médio da variável independente e é dado pela expressão 4.10.

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4.10}$$

Esse coeficiente é uma forma de mensurar o quanto a função regressora se ajusta aos dados. Expressa a proporção da variância da estimativa da variável dependente em torno de sua média, e deve assumir valores entre 0 e 1, de modo que quanto mais próximo de 1, mais representativa é a predição (HAIR et al., 2009).

As métricas definidas pelas equações 4.5, 4.6, 4.8 e 4.9 são utilizadas nesse trabalho para avaliar individualmente cada um dos atributos estimados, o deslocamento em relação ao eixo x e ao eixo y, no plano cartesiano e os tamanhos dos eixos da elipse de tempestade.

Contudo, para obter uma avaliação mais expressiva quanto à previsão do deslocamento, é adotada a relação entre o centróide observado da tempestade e o formado pela previsão do deslocamento em cada eixo.

Nesse sentido, sejam C_i a posição identificada do centróide e E_i a estimada, de modo que i representa a i-ésima tempestade, a distância entre os pontos é dada por:

$$d_i = \mid \overrightarrow{C_i E_i} \mid . \tag{4.11}$$

Dessa forma, a partir da distância para cada célula do conjunto de dados, duas métricas são definidas para avaliar os modelos, conforme apresentadas pelas equações 4.12 e 4.13.

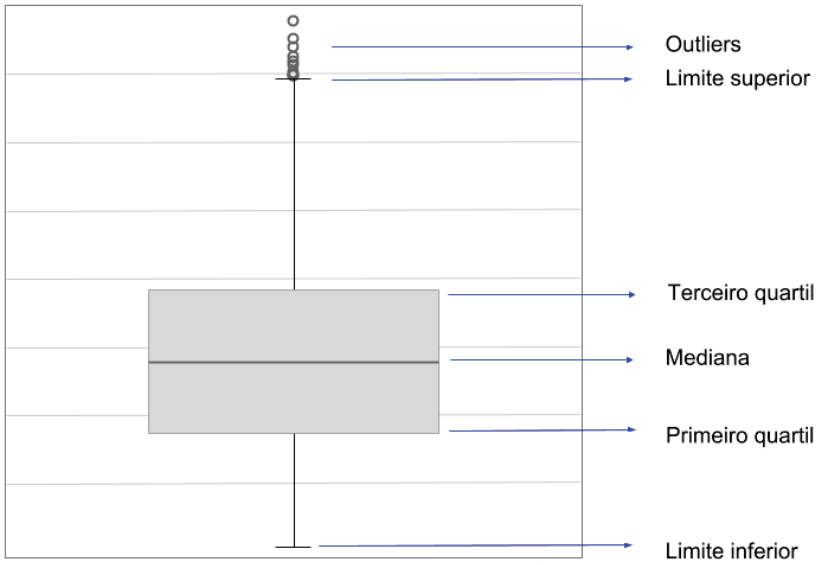
$$\mu = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{4.12}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \mu)^2} \tag{4.13}$$

A primeira trata da média aritmética das distâncias ou erro médio, e permite uma visão geral do erro para o conjunto de dados. Já a segunda, representa o desvio padrão, ou seja, mostra a dispersão das distâncias indiviuais em relação ao erro médio.

Uma forma de visualizar a dispersão e identificar assimetrias em dados é o gráfico *boxplot*, que baseia-se na divisão em quartis, conforme mostra a FIGURA 20. Essa partição premite determinar o *desvio-quarti*, que é dado pela diferença entre o terceiro e o primeiro quartis, representado no gráfico como a região retangular (SILVA, 2009).

FIGURA 20 – ESTRUTURA DE UM BOXPLOT



FONTE: A autora (2018)

Os limites inferior e superior são determinados tomando-se uma proporção de 1,5 da *amplitude iterqualítica* acima do terceiro quartil para o superior e abaixo do primeiro para o inferior. Quando o desvio-quartil leva a valores fora dos limites do conjunto, os valores mínimo e máximo dos dados são tomados. Essa métrica permite identificar os dados denominados *outliers* ou ruidosos, que são aqueles que excedem os limites tomados, mostrando uma grande discrepância da distribuição da amostra.

Essa forma de visualização dos dados é bastante utilizada. Permite a avaliação de diversas características da amostra, desde a simples mediana, à dispersão e ruídos presentes nos dados.

4.2.2 Seleção das características

Um conjunto de diferentes tipos de características foi gerado neste trabalho. Contudo, não há conhecimento prévio da relevância de cada atributo para a previsão proposta. Em outras palavras, o uso dos dados disponíveis diretamente nos modelos pode incluir características com um nível de redundância relativamente grande, o que poderia ser corrigido eliminando-se aquelas que são altamente correlacionadas.

Em contrapartida, tal metodologia não permitiria encontrar quais características proporcionam um melhor aprendizado. Objetiva-se não apenas eliminar atributos que, por apresentarem grande similaridade, não contribuem para a melhoria do modelo, mas também aqueles que não são representativos aos fins em que os modelos se aplicam.

Com esse intuito, a seleção recursiva de características é utilizada, apresentada por Guyon et al. (2002), que baseia-se no conhecimento do peso de cada um dos atributos. Nos métodos de modelo linear esses pesos são dados pelos coeficientes estimados. Já para os *ensembles*, são calculados baseados na importância de Gini, que toma sobre todos os nós e todas as árvores a soma da redução da impureza alcançada, assumindo que as variáveis mais importantes são aquelas que mais reduzem a impureza, como mostra Louppe et al. (2013).

O peso de cada variável é dado atribuindo-se um valor entre zero e um, por meio da execução do algoritmo de aprendizado, e são ordenados em um *ranking*. A ordenação permite que os recursos de menor importância sejam excluídos, um a cada iteração, em que são recalculados para cada subconjunto menor. Avalia-se a função de erro e o processo iterativo é executado até que não ocorra mais melhoras no desempenho do algoritmo.

Desse modo, o processo recursivo é executado para cada uma das técnicas, em que seus subconjuntos ótimos de atributos são armazenados. Cada um dos subconjuntos é testado para todos os algoritmos, visto que a exclusão de um elemento por vez, recursivamente, pode levar a um mínimo local.

A partir destas execuções encontrou-se um conjunto que apresenta os melhores resultados para todos os modelos utilizados na pesquisa, tomado como o padrão para todo o trabalho, como mostra o QUADRO 8.

Nesta tabela, o parâmetro $T - 20$ representa o dado histórico de vinte minutos antecedente ao tempo inicial de previsão, $\Delta(T - 20, T - 10)$ e $\Delta(T - 10, T)$ representam a diferença, ou variação, dos atributos em relação a 20 e 10 minutos antecessores e, analogamente, de 10 para o inicial.

Além das características históricas apresentadas, duas outras são utilizadas: o mês e horário de ocorrência, para fazer menção ao período do ano e do dia em que

QUADRO 8 – CONJUNTO DE ATRIBUTOS

Histórico	**Características**
$T - 20$	Orientação da elipse, eixo menor e maior da elipse, velocidade de deslocamento, direção do deslocamento, área, VIL, números total de raios, x, y.
$\Delta(T - 20, T - 10)$	Orientação da elipse, eixo menor e maior da elipse, velocidade de deslocamento, direção do deslocamento, área, VIL, número de raios positivos, números total de raios, x, y.
$\Delta(T - 10, T)$	Orientação da elipse, eixo menor e maior da elipse, velocidade de deslocamento, direção do deslocamento, área, VIL,número de raios negativos e positivos, números total de raios, x, y.

FONTE: A autora (2018)

cada tempestade ocorreu.

4.2.3 Especificações dos métodos

Um método amplamente conhecido de validação cruzada, o k-*fold*, foi utilizado na escolha dos parâmetros. Basicamente, consiste na divisão do conjunto de dados em k subamostras de elementos mutuamente exclusivos, com as quais o algoritmo é executado k vezes e em cada momento uma das subamostras é utilizada para teste e as $k - 1$ restantes formam o conjunto de treinamento, como pode ser encontrado em Mohri, Rostamizadeh e Talwalkar (2012).

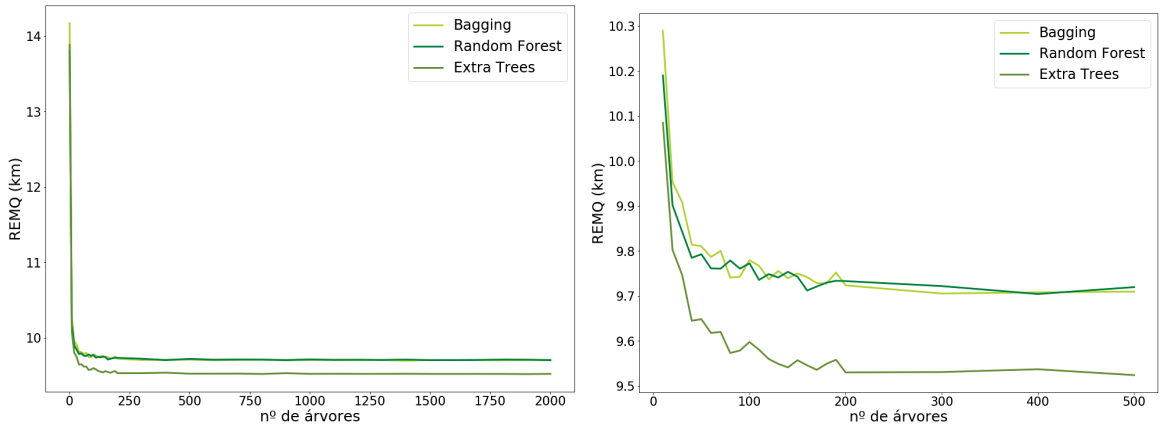
O erro da validação cruzada é dado pela média dos k erros encontrados. A métrica assumida para esse processo é a raiz do erro médio quadrático, dada pela equação 4.8.

Nos métodos de aprendizagem agrupada, um importante parâmetro é o número de estimadores, que deve ser estabelecido de forma que obtenha-se a maior variedade possível no conjunto. Assim, estabelecido o método 3-*fold*, as REMQ da validação cruzada para os diferentes números de árvores avaliados são mostradas na FIGURA 21.

Em relação ao *Random Forest*, Breiman (2001) afirma que "o erro de generalização para florestas converge a um limite quando o número de árvores na floresta se torna grande", também é possível ver isso para o *Bagging* e *Extra Trees*, no gráfico apresentado na FIGURA 21.

Portanto, duzentas árvores é um número suficiente para os três métodos avaliados. As técnicas *Random Forest* e *Extra Trees* têm outros dois parâmetros importantes que são o número máximo de características a serem avaliadas na seleção de cada nó, e a máxima profundidade de cada árvore.

FIGURA 21 – NÚMERO DE ÁRVORES PARA AS TÉCNICAS DE ENSEMBLE

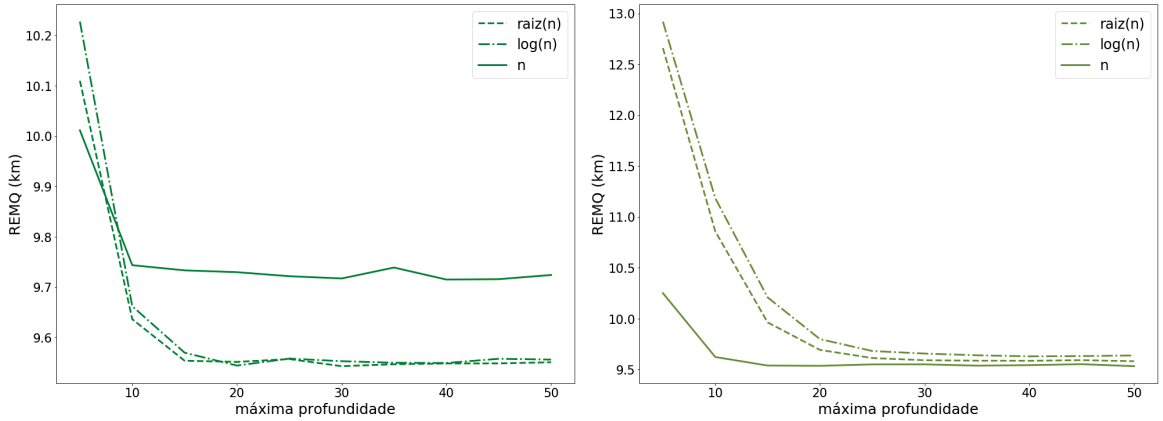


(a) 1 à 2000 árvores (b) 10 à 500 árvores

FONTE: A autora (2018)

Três variações são avaliadas, considerando-se n o número total de exemplos na amostra de treinamento, verifica-se o uso da raiz quadrada de n, seu logarítmo e, ainda, o número total de exemplos, cada uma delas com a profundidade das árvores de cinco a cinquenta, como mostra a FIGURA 22.

FIGURA 22 – MÁXIMA PROFUNDIDADE DAS ÁRVORES E DE NÚMERO DE ATRIBUTOS PARA AS TÉCNICAS DE ENSEMBLE



(a) Random Forest (b) Extra Trees

FONTE: A autora (2018)

Com isso, foram definidas a profundidade máxima igual a quinze para ambos os algoritmos, enquanto o número máximo de atributos como a raiz de n para o método *Random Forest* e o número total de amostras para o *Extra Trees*, como é apresentado na TABELA 4.

Quanto aos métodos de modelo linear, para o *Bayesian Ridge* estabeleceu-se um máximo de 5000 iterações e a tolerância de 1×10^{-9}, como também nos demais métodos.

TABELA 4 – MÁXIMA PROFUNDIDADE DAS ÁRVORES DE DECISÃO (ENSEMBLES)

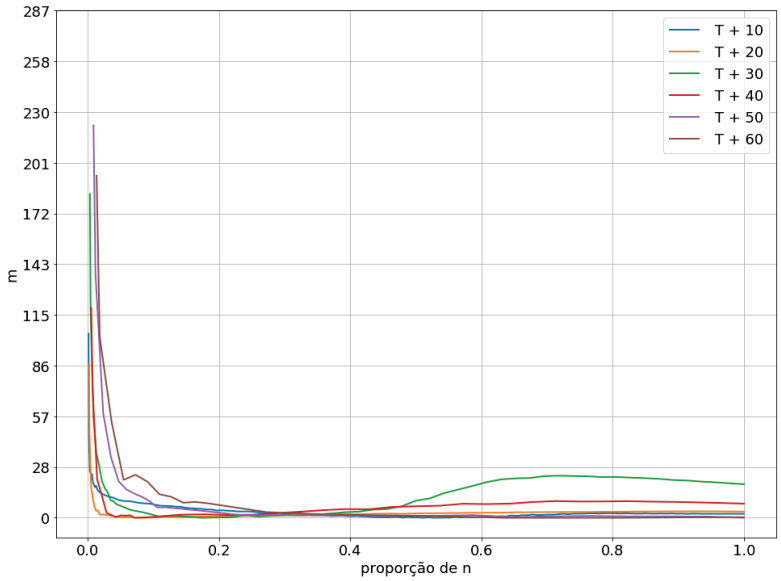
Tempo de previsão	Random Forest	Extra Trees
T + 10 minutos	201	40407
T + 20 minutos	157	24789
T + 30 minutos	122	14774
T + 40 minutos	94	8881
T + 50 minutos	75	5635
T + 60 minutos	60	3592

FONTE: A autora (2018)

No método *Theil Sen*, a escolha de m para a combinação simples de m em n é analisada, a partir de 5-*fold*. Uma análise em relação ao número de exemplos é realizada, iniciando com o valor mínimo possível $k + 1$, onde k representa o número de atributos, e tomando valores até obter número total de exemplos de treinamento, dado por n. Como estabelecido na subseção 3.4.1, esses valores são os limites para o parâmetro.

Como o conjunto de dados para cada período de previsão é formado por um número distinto de exemplos, toma-se, por padronização, uma resolução percentual do número de amostras para a análise, variando de $k + 1$, seguido das proporções (5%, 10%, ..., 100%) de n, como apresentado na FIGURA 23. No eixo vertical é dada a variação do raiz do erro médio quadrático, em metros. Denota-se por zero o menor erro avaliado para cada período de previsão, ou seja, o melhor ajuste encontrado. Os demais valores, avaliados no eixo vertical do gráfico, referem-se a variação do desempenho de cada um dos modelos avaliados em relação ao melhor encontrado, mensurada em metros.

FIGURA 23 – ESCOLHA DO PARÂMETRO m COMO PROPORÇÃO DE n



FONTE: A autora (2018)

Pode-se observar a partir dos valores do parâmetro m, que existe um ponto, próximo da proporção de 30%, que a maioria dos métodos se encontram e apresentam um valor bem próximo de zero, ou seja, do melhor modelo encontrado, em que a maior variação de desempenho não ultrapassa 3,22 metros, em relação ao menor valor que poderia assumir, ou seja, apresenta uma variação bem pequena. Com base nisso, toma-se m como 28% de n. A TABELA 5 apresenta os valores numéricos do parâmetro para cada período de previsão, visto que o número de tempestades utilizadas varia de acordo com o tempo que deseja-se prever.

TABELA 5 – PARÂMETRO m PARA CADA PERÍODO DE PREVISÃO

Tempo de previsão	**parâmetro** m
T + 10 minutos	11314
T + 20 minutos	6941
T + 30 minutos	4137
T + 40 minutos	2487
T + 50 minutos	1578
T + 60 minutos	1006

FONTE: A autora (2018)

Esse resultado pode ser avaliado do ponto de vista que, segundo a literatura, quanto menor o valor de m em relação ao mínimo e máximo possível, mais robusto o algoritmo se torna. A partir disso, é possível verificar que a mediana multivariada toma uma proporção mais voltada à robustez, do que à eficiência ou a regressão do próprio MMQ.

Nesse capítulo, foram estabelecidos os conjuntos de aprendizado, a manipulação dos dados e as especificações para geração de cada um dos modelos, a partir dos algoritmos de AM. As seguintes siglas foram usadas para denominação dos modelos estimadores: RFE (*Random Forest Estimator*), ETE (*Extra Trees Estimator*), TBE (*Tree-Bagging Estimator*), BRE (*Bayesian Ridge Estimator*) e TSE (*Theil Sen Estimator*).

5 RESULTADOS

Os modelos, desenvolvidos conforme a seção 4.2, têm seus resultados apresentados em quatro seções: a primeira (5.1) trata da validação, como estabelecida na subseção 4.1.3. A segunda e terceira (5.2 e 5.3) apresentam cada um dos grupos de AM utilizados. Por fim, a quarta (5.4) se dedica a avaliação dos métodos de aprendizado comparados com o TITAN.

5.1 VALIDAÇÃO DO MODELO

Para efeitos de validação, considera-se a previsão do deslocamento e analisa-se a distância do ponto previsto ao observado pelas métricas de média e desvio padrão. Nesse sentido, a média das distâncias para todos os métodos e períodos de previsão são apresentadas na FIGURA 24 em quilômetros, os modelos de *ensemble* são representados pelas diferentes tonalidades da cor verde, enquanto os de modelo linear azul e o TITAN preto.

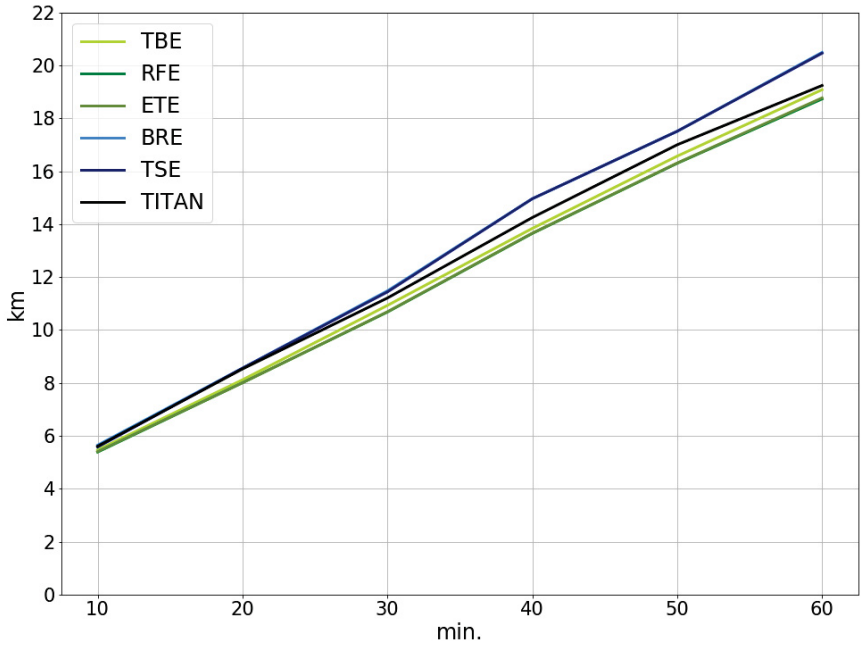
É possível visualizar que as distâncias médias foram maiores no conjunto de validação para todos os métodos, inclusive para o TITAN. Apesar desse aumento, o mesmo padrão pode ser observado nos gráficos. Os métodos de *ensemble*, apesar de apresentarem uma diferença maior na validação, se mantêm com um desempenho melhor. Enquanto isso, os métodos de modelo linear que coincidiam no conjunto de teste, o TSE (*Theil Sen Estimator*) se mantém pior com aumento dos erros em relação ao TITAN, e o BRE (*Bayesian Ridge Estimator*) passa a reduzir a média mostrando resultados bem semelhantes ao do software.

Quanto ao desvio padrão, de uma forma geral, os resultados são similares às distâncias médias, como pode ser visualizado na FIGURA 25. Em oposição à média, todos os métodos apresentam uma redução do desvio na validação, em relação ao conjunto de teste.

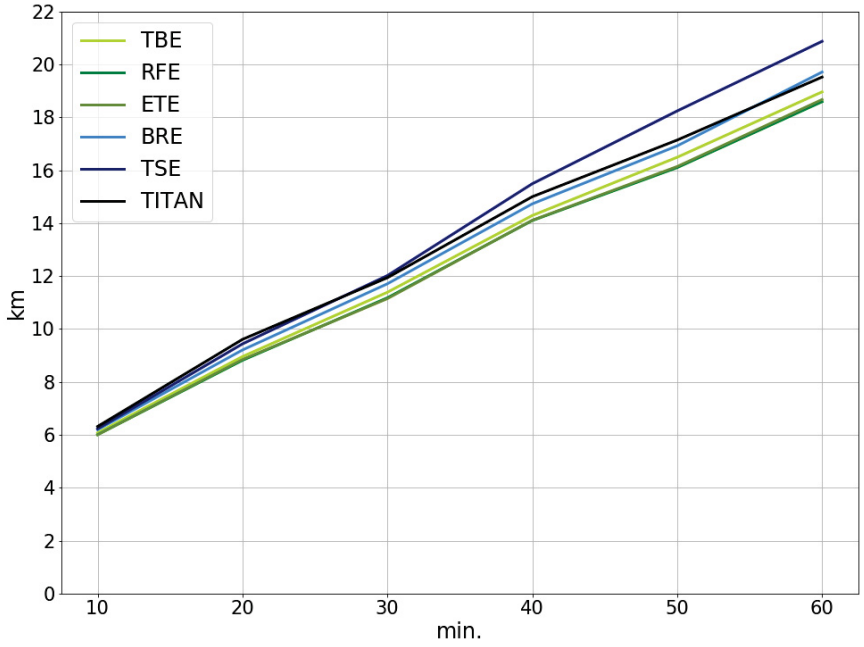
Nessa segunda métrica, todos os métodos de modelo linear apresentam melhores resultados para a validação e, ainda, os algoritmos de AM apresentam uma redução maior quando comparados ao TITAN.

Com isso, verifica-se que para um conjunto totalmente externo ao conjunto de dados geral, os resultados se mantêm e até mesmo apresentam ganhos. Isso indica que a generalização do conjunto de teste não apresentou indução por ser escolhido aleatoriamente, ou seja, indifere da distância temporal das células.

FIGURA 24 – MÉDIA DAS DISTÂNCIAS DOS CENTRÓIDES PREVISTOS AOS OBSERVADOS



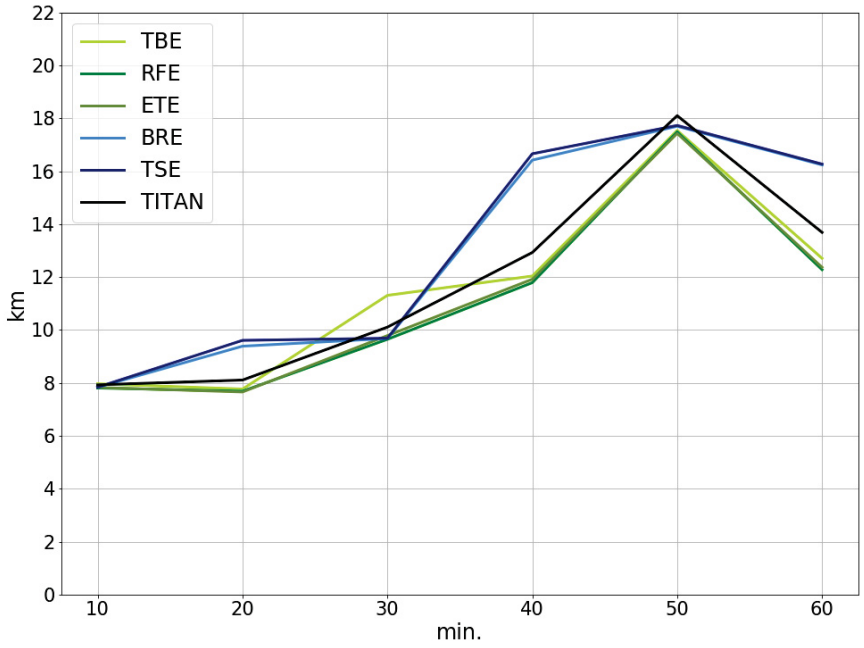
(a) Conjunto de teste



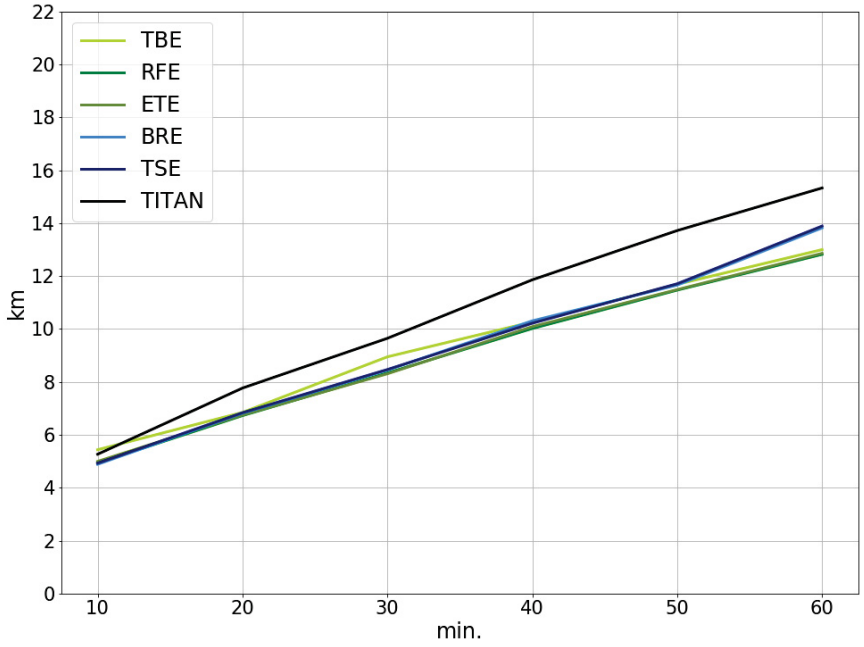
(b) Conjunto de validação

FONTE: A autora (2018)

FIGURA 25 – DESVIO PADRÃO DAS DISTÂNCIAS DOS CENTRÓIDES PREVISTOS AOS OBSERVADOS



(a) Conjunto de teste



(b) Conjunto de validação

FONTE: A autora (2018)

Dessa forma, a partir desse momento o conjunto de validação é tomado para a análise dos demais resultados, pois representa um número maior de dados.

5.2 MÉTODOS DE ENSEMBLE

Os métodos de *ensemble* apresentam desempenhos bem semelhantes entre si, como pode ser visualizado pelas FIGURAS 24 e 25, onde os valores destacados em negrito representam os melhores resultados para cada período. Quanto a distância média, as técnicas *Random Forest* (RFE) e *Extra Trees* (ETE) diferem em menos de 30 metros em cada um dos períodos de previsão, sendo que apenas para 30 minutos o ETE apresenta o melhor resultado, como apresentado na TABELA 6.

TABELA 6 – MÉDIA DAS DISTÂNCIAS DO PONTO PREVISTO AO OBSERVADO EM QUILÔMETROS (*ENSEMBLES*)

Método	**T + 10**	**T + 20**	**T + 30**	**T + 40**	**T + 50**	**T + 60**
TBE	6,083	8,962	11,388	14,295	16,485	18,96
RFE	**6,006**	**8,82**	11,18	**14,101**	**16,102**	**18,588**
ETE	6,013	8,845	**11,149**	14,109	16,14	18,669

FONTE: A autora (2018)

O algoritmo *Bagging* (TBE) fica um pouco atrás dos outros em desempenho, podendo diferir em quase 400 metros, como pode ser verificado para a previsão de 50 minutos, por exemplo.

Essas interpretações se mantêm em relação ao desvio padrão, com um único adendo de que a diferença entre os métodos nessa métrica é ainda menor, conforme apresenta a TABELA 7.

TABELA 7 – DESVIO PADRÃO DAS DISTÂNCIAS DO PONTO PREVISTO AO OBSERVADO EM QUILÔMETROS (*ENSEMBLES*)

Método	**T + 10**	**T + 20**	**T + 30**	**T + 40**	**T + 50**	**T + 60**
TBE	5,443	6,848	8,949	10,272	11,686	13,002
RFE	4,942	**6,726**	8,338	**10,024**	**11,47**	**12,823**
ETE	**5,003**	6,752	**8,308**	10,093	11,488	12,863

FONTE: A autora (2018)

Os algoritmos *Random Forest* e *Extra Trees* que implementam mais aleatoriedade do que o *Bagging* apresentam melhores resultados, o que é coerente com a revisão teórica. As duas técnicas obtiveram desempenhos muito similares, em que

destacou-se o RFE.

5.3 MÉTODOS DE MODELO LINEAR

O método *Bayesian Ridge* (BRE) destaca-se no grupo de modelo linear, podendo superar o desempenho de *Theil Sen* (TSE) em mais de um quilômetro, como mostra a TABELA 8.

TABELA 8 – MÉDIA DAS DISTÂNCIAS DO PONTO PREVISTO AO OBSERVADO EM QUILÔMETROS (MODELOS LINERES)

Método	**T + 10**	**T + 20**	**T + 30**	**T + 40**	**T + 50**	**T + 60**
BRE	**6,204**	**9,206**	**11,709**	**14,726**	**16,925**	**19,714**
TSE	6,232	9,435	12,023	15,49	18,243	20,867

FONTE: A autora (2018)

O desvio padrão das distâncias apresenta números bem semelhantes entre si, no qual TSE supera em noventa metros o BRE para a previsão de 40 minutos, como mostra a TABELA 9.

TABELA 9 – DESVIO PADRÃO DAS DISTÂNCIAS DO PONTO PREVISTO AO OBSERVADO EM QUILÔMETROS (MODELOS LINEARES)

Método	**T + 10**	**T + 20**	**T + 30**	**T + 40**	**T + 50**	**T + 60**
BRE	**4,89**	**6,822**	**8,445**	10,312	**11,664**	**13,821**
TSE	4,928	6,836	8,469	**10,222**	11,71	13,888

FONTE: A autora (2018)

Contudo, a análise dos resultados quanto ao centróide previsto, com base na previsão do deslocamento em relação aos eixos x e y, leva a concluir que o método baseado no teorema de Bayes se adaptou melhor ao problema do que o de mediana multivariada.

5.4 APRENDIZADO DE MÁQUINA E O TITAN

A distância do centróide previsto ao observado, permite avaliar a previsão do deslocamento, principal objetivo desta pesquisa. Esses resultados são avaliados para cada um dos grupos de métodos e, agora, visando avaliar a qualidade deles,

toma-se o desempenho do TITAN, para o mesmo conjunto de dados, conforme mostra a TABELA 10.

TABELA 10 – AVALIAÇÃO DAS DISTÂNCIAS DO PONTO PREVISTO AO OBSERVADO EM QUILÔMETROS (TITAN)

Métrica	**T + 10**	**T + 20**	**T + 30**	**T + 40**	**T + 50**	**T + 60**
Média	6,324	9,612	11,942	15,002	17,138	19,524
Desvio Padrão	5,267	7,774	9,654	11,864	13,719	15,33

FONTE: A autora (2018)

As FIGURAS 24 e 25 apresentadas com objetivo de validação já dão algumas interpretações quanto ao desempenho de cada método e de cada grupo. Para incorporar a verificação dos resultados, a FIGURA 26 permite ter uma visão geral sobre a distribuição das distâncias calculadas a partir da estimativa de cada método. A mediana é representada pela linha de cor laranja, e o valor médio pelo losango branco.

É possível verificar que, em relação ao TITAN, as técnicas de AM apresentam um ganho no desempenho tanto em relação a média, quanto a dispersão geral dos dados, com excessão do TSE. A mediana, por outro lado, se mostra menor para o TITAN, juntamente com os *ensembles* ETE e RFE, diferindo pouco entre si. Na estimativa de 60 minutos, o software supera todas as técnicas de aprendizado nessa métrica.

O centróide previsto é obtido por meio da posição conhecida para T, período inicial da previsão, e estimando-se o deslocamento em relação ao norte (eixo y) e ao leste (eixo x) para cada uma da células. O processo de Aprendizado de Máquina se deu em cada um desses deslocamentos individuais, que podem ser avaliados pelo coeficiente de determinação, como mostram as TABELAS 11 e 12.

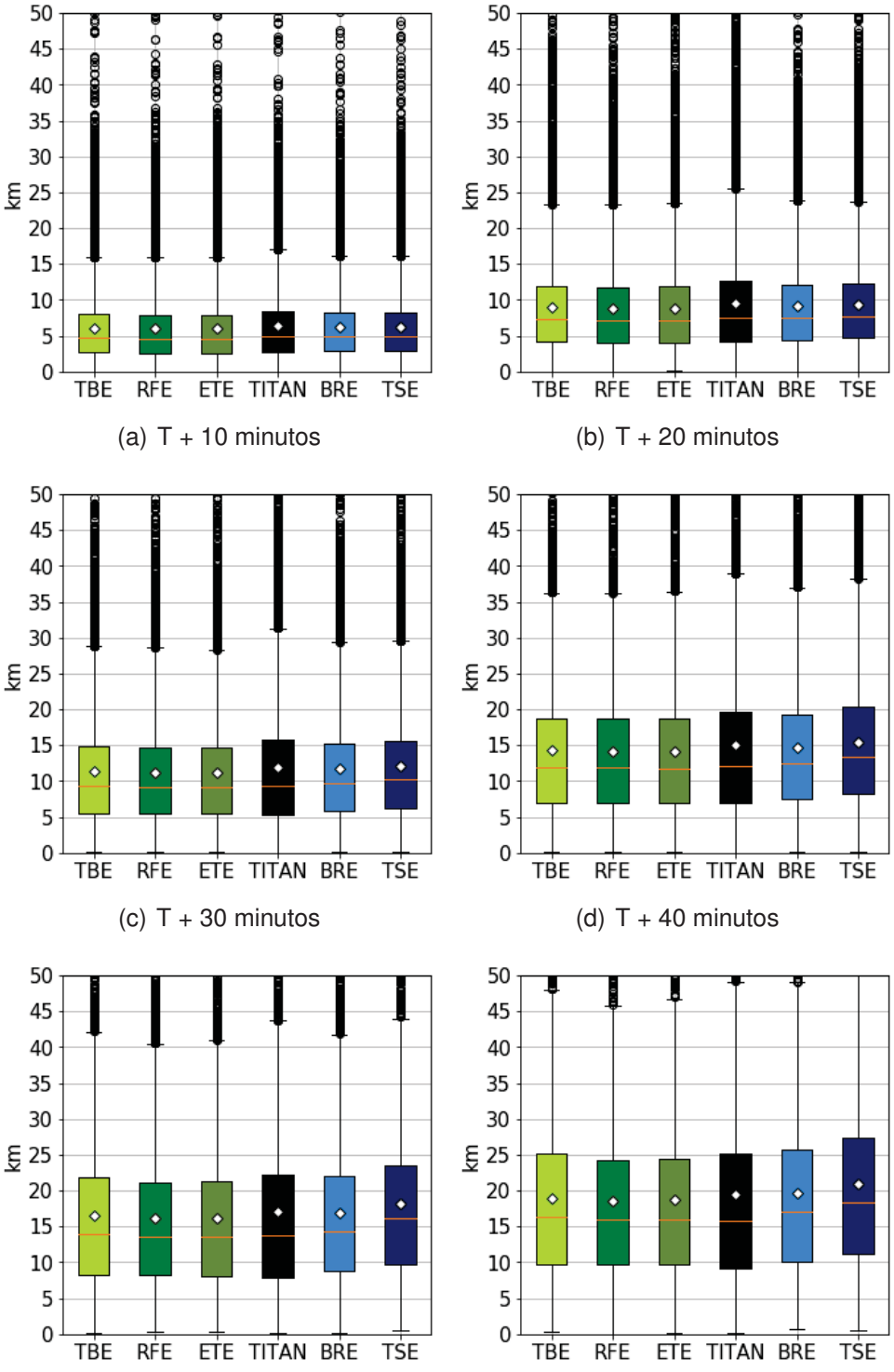
TABELA 11 – COEFICIENTE DE DETERMINAÇÃO (PREVISÃO DO DESLOCAMENTO NO EIXO X)

Método	**T + 10**	**T + 20**	**T + 30**	**T + 40**	**T + 50**	**T + 60**
TBE	0,319	0,483	0,604	0,625	0,674	0,713
RFE	**0,335**	0,498	**0,617**	**0,641**	**0,692**	0,723
ETE	0,333	**0,499**	**0,617**	**0,641**	**0,692**	**0,728**
BRE	0,306	0,467	0,586	0,617	0,674	0,699
TSE	0,296	0,429	0,554	0,577	0,603	0,654
TITAN	0,237	0,371	0,523	0,551	0,614	0,67

FONTE: A autora (2018)

Como discutido na subseção 4.2.1, quanto mais próximo de 1, melhor a estimativa do método. Com isso, pode-se verificar que, quanto ao deslocamento horizontal, as métricas de AM apresentam melhor ajuste, com excessão apenas de *Theil Sen* na estimativa para 50 e 60 minutos.

FIGURA 26 – BOXPLOT DA DISTRIBUIÇÃO DA DISTÂNCIA DO PONTO PREVISTO AO OBSERVADO



(a) T + 10 minutos

(b) T + 20 minutos

(c) T + 30 minutos

(d) T + 40 minutos

(e) T + 50 minutos

(f) T + 60 minutos

FONTE: A autora (2018)

TABELA 12 – COEFICIENTE DE DETERMINAÇÃO (PREVISÃO DO DESLOCAMENTO NO EIXO Y)

Método	T + 10	T + 20	T + 30	T + 40	T + 50	T + 60
TBE	0,114	0,288	0,426	0,412	**0,529**	**0,547**
RFE	**0,184**	**0,343**	**0,438**	**0,486**	0,524	0,545
ETE	0,168	0,34	0,435	0,483	0,525	0,542
BRE	0,117	0,298	0,393	0,441	0,48	0,472
TSE	0,145	0,273	0,341	0,361	0,466	0,411
TITAN	0,109	0,195	0,326	0,39	0,427	0,453

FONTE: A autora (2018)

Do mesmo modo acontece para o deslocamento vertical, em que TSE apresenta resultados inferiores para 40 e 60 minutos de previsão. As demais técnicas se mantêm com um desempenho superior.

A partir das previsões do deslocamento e conhecida a informação da posição do centróide da tempestade no período T, as coordenadas x e y do novo centróide são geradas. Quanto a essas cordenadas, a raiz do erro médio quadrático e o erro médio absoluto são avaliadas, como apresentado pela FIGURA 27.

Tanto para a coordenada x quanto para a y, os algoritmos de AM, com excessão do TSE, apresentam o EMA com valores menores ao do TITAN, assim como o REMQ. Além disso, em vários períodos, as avaliações das duas métricas se encontram mais próximas para esses métodos, o que mostra que além do desempenho médio ser melhor, a dispersão em torno desse valor também é menor.

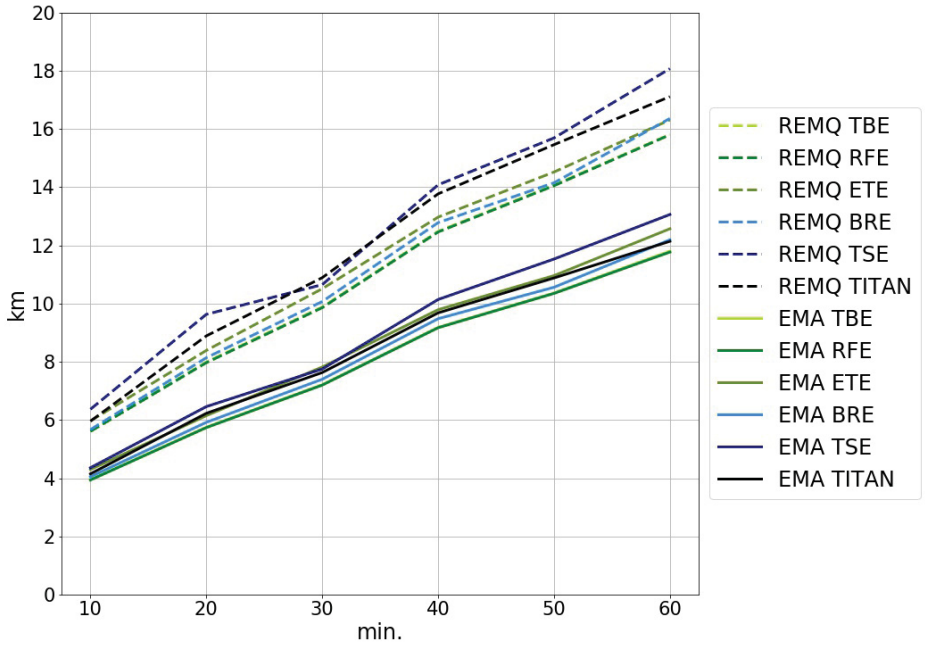
Quando avaliado o EMA, *Theil Sen* apresenta resultados inferiores aos obtidos pelo software TITAN, podendo superá-lo em algumas posições. Quanto ao REMQ, o método o supera para a maioria dos períodos, ou seja, apesar de mostrar uma estimativa inferior, permite uma menor discrepância entre os valores estimados.

Essas previsões permitem a incorporação da estimativa da velocidade do deslocamento do centróide. A avaliação do desempenho pode ser obtida diretamente proporcional ao a estimativa do deslocamento, levando-se em consideração a resolução temporal de 10 minutos.

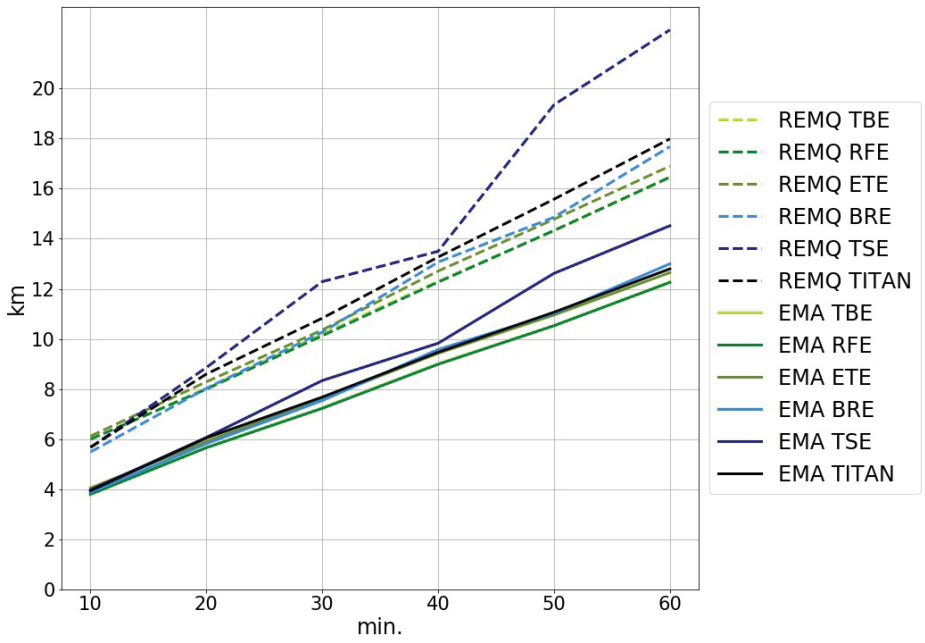
Visto que os algoritmos de aprendizado mostraram bons resultados para o problema, os mesmos modelos podem ser usados com as características e parâmetros já selecionados, para a previsão dos tamanhos dos eixos da tempestade. Esse processo se dá realizando o treinamento para as mesmas amostras, mas com as novas variáveis como valores desejados. Em termos de erro médio absoluto e a raiz quadrática, os resultados encontrados são apresentados na FIGURA 28.

O desempenho de todos os métodos de AM destaca-se, tanto para o eixo maior das células, quanto o menor. Além do erro médio absoluto apresentar valores

FIGURA 27 – ERRO DE PREVISÃO DO DESLOCAMENTO



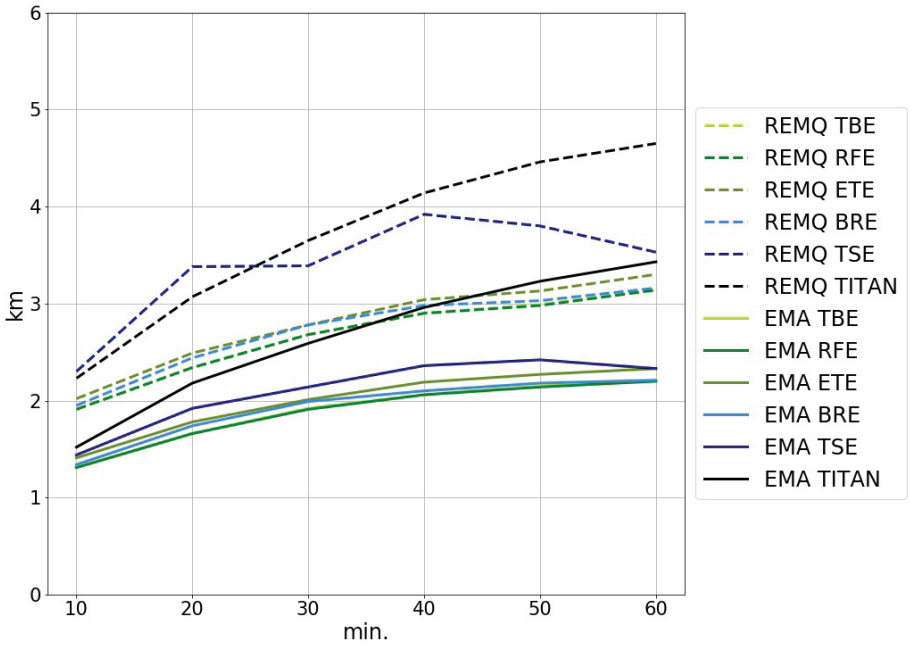
(a) Eixo x



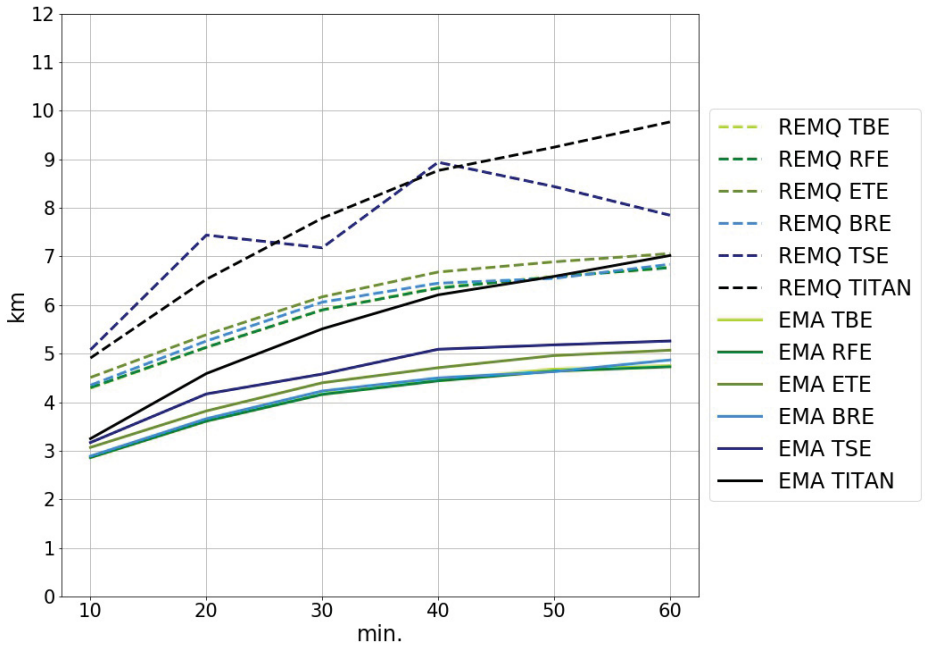
(b) Eixo y

FONTE: A autora (2018)

FIGURA 28 – ERRO DE PREVISÃO DO TAMANHO DOS EIXOS



(a) Eixo menor



(b) Eixo maior

FONTE: A autora (2018)

mais baixos em todas as posições, para os períodos de 50 e 60 minutos, a maior parte deles supera a raiz do erro médio quadrático do software. Ou seja, para esses momentos, mesmo considerando toda a dispersão das estimativas para os algoritmos de aprendizado, o erro médio do TITAN apresenta resultados inferiores.

Outro ponto interessante, é a diferença entre o REMQ e EMA que, assim como nas interpretações para o deslocamento, se mantêm menores aos obtidos pelo TITAN, indicando uma menor discrepância entre as estimativas.

Com base nessas previsões é possível estimar a área (A) da tempestade, por meio da equação 5.1, em que a e b representam o tamanho do semieixo maior e menor da elipse, respectivamente, como dado na FIGURA 5.

$$A = \pi ab \tag{5.1}$$

Por meio do coeficiente de determinação, o cômputo pode ser examinado pela TABELA 13.

TABELA 13 – COEFICIENTE DE DETERMINAÇÃO (ÁREA ESTIMADA DA ELIPSE)

Método	**T + 10**	**T + 20**	**T + 30**	**T + 40**	**T + 50**	**T + 60**
TBE	0,935	0,894	**0,777**	0,681	0,436	0,309
RFE	**0,936**	**0,895**	0,776	**0,689**	**0,456**	**0,354**
ETE	**0,936**	0,894	0,771	0,674	0,443	0,347
BRE	0,928	0,849	0,715	0,489	0,411	0,291
TSE	0,927	0,845	0,705	0,457	0,386	0,203
TITAN	0,915	0,809	0,557	0,367	-0,381	-0,633

FONTE: A autora (2018)

O decréscimo da acurácia das predições para todos os algoritmos, de acordo com o aumento do tempo, é evidente. Para o TITAN mostra-se ainda mais acentuado.

Como já foi verificado pela mensuração dada na FIGURA 28, o destaque das abordagens de AM é perceptível. Com o R^2 da área, isso torna-se, visivelmente, destoante, apresentando valores negativos para a técnica nas últimas duas posições. Definiu-se que o valor do coeficiente é limitado entre zero e um. Então, se faz necessária a análise da métrica, para o caso em que seu valor é negativo.

Tem-se na equação 4.9 o R^2 definido pela diferença entre 1 e a expressão apresentada a seguir (5.2). Assim, só é possível que R^2 assuma um valor inferior a zero, se essa fração tomar um valor superior a 1.

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5.2}$$

Como definido na subseção 4.2.1, $\overline{y}$ representa o valor médio dos rótulos conhecidos e $\hat{y}_i$ o valor estimado para o i-ésimo exemplo da amostra.

Em análise da expressão apresentada, o numerador e o denominador da fração diferem unicamente em relação ao termo em que se calcula diferença. O denominador é dado pela soma do erro quadrático (SEQ), para o caso em que $\hat{y}$ representa os valores estimados.

Nesse sentido, se for tomado simplesmente o valor médio esperado, como a estimativa para qualquer y, pode-se dizer que o numerador representa a mesma métrica, para este caso. Em outras palavras, observa-se que o coeficiente de determinação, só deve assumir um valor negativo se o SEQ do modelo avaliado for maior do que quando toma-se a média como hipótese. Por esse motivo, avalia-se R^2, nos limites de 0 a 1, para regressão, em que o seu mínimo implica que o valor médio possui igual ajuste aos dados.

Nesse contexto, para períodos maiores que 40 minutos, a predição do software TITAN se faz insuficiente para a área. Mesmo para essas posições, apesar do decréscimo no desempenho, os modelos de AM se mantêm com uma acurácia aceitável e se apresentam como uma ótima proposta para substituição do modelo padrão.

Após avaliadas as diferentes métricas para as predições permitidas pelos métodos, um caso para visualização é selecionado.

Os critérios para a seleção se baseiam que o evento deve pertencer ao conjunto de validação e que a variação da distância entre o método de AM que apresentou melhor resultado, o *Random Forest*, e o TITAN deve ser condizente com os resultados apresentados.

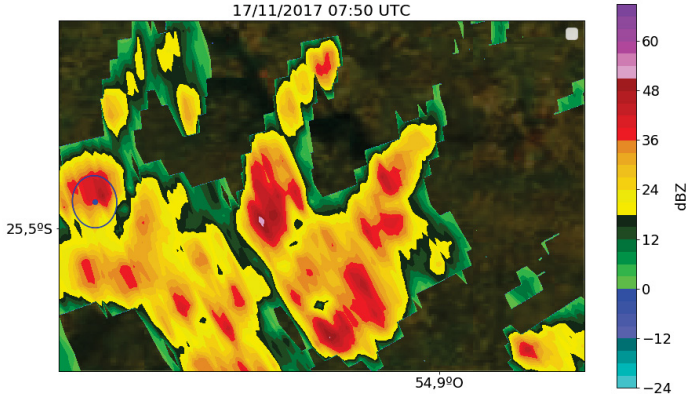
Além disso, a tempestade deve ter ocorrido, desde a sua primeira identificação até a sua máxima previsão, 60 minutos, na região de um dos radares meteorológicos do SIMEPAR, localizado em Cascável, no Paraná.

O evento selecionado ocorreu em 17 de novembro de 2017, com a sua primeira identificação as 7 horas e 50 minutos, em tempo coordenado universal (UTC), como apresenta a FIGURA 29.

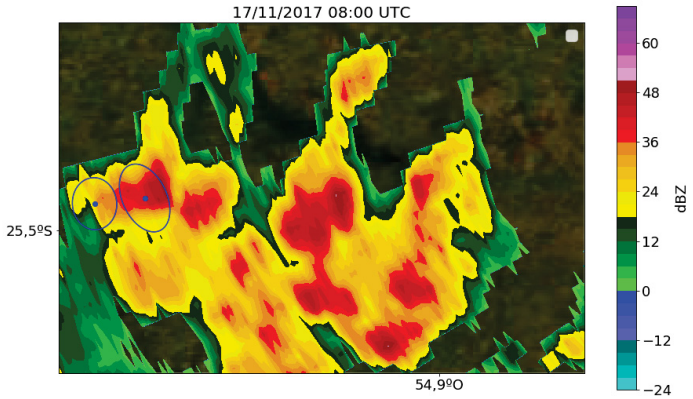
Assim que o seu terceiro centróide é identificado, as técnicas realizam a previsão para as 6 posições subsequentes e geram a rota prevista de deslocamento, como mostram a FIGURA 30, FIGURA 31 e FIGURA 32 para os métodos de aprendizado agrupado, respectivamente, *Bagging*, *Random Forest* e *Extra Trees* e a FIGURA 33 e FIGURA 34 para os de modelo linear, respectivamente, *Bayesian Ridge* e *Theil Sen*.

O caso selecionado permite visualizar que, apesar de todas as rotas preverem um deslocamento maior que o observado, os *ensembles* descrevem melhor o movi-

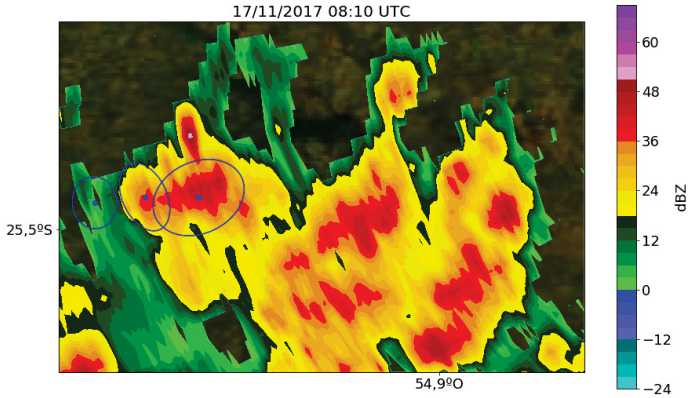
FIGURA 29 – PRIMEIRAS IDENTIFICAÇÕES (17/11/2017)



(a) T - 20 minutos



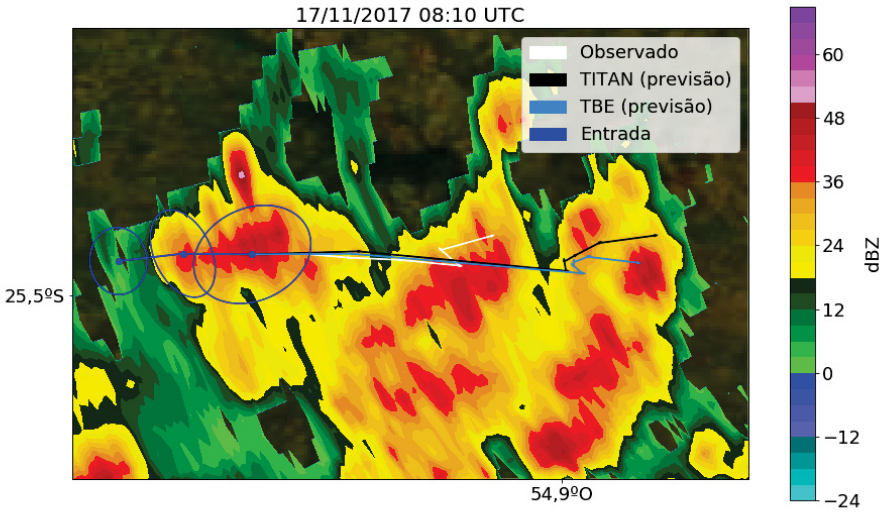
(b) T - 10 minutos



(c) T + 0 minuto

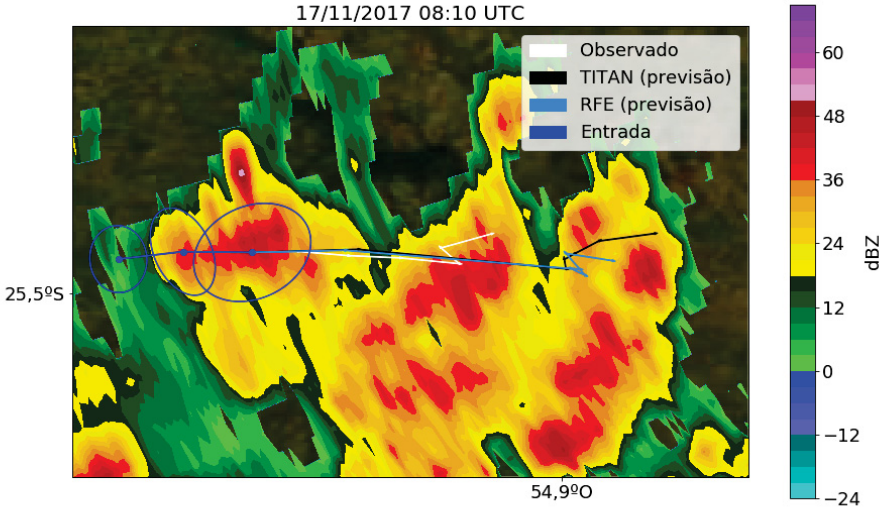
FONTE: A autora (2018)

FIGURA 30 – PREDIÇÃO DO MÉTODO BAGGING (17/11/2017 8:10 UTC)



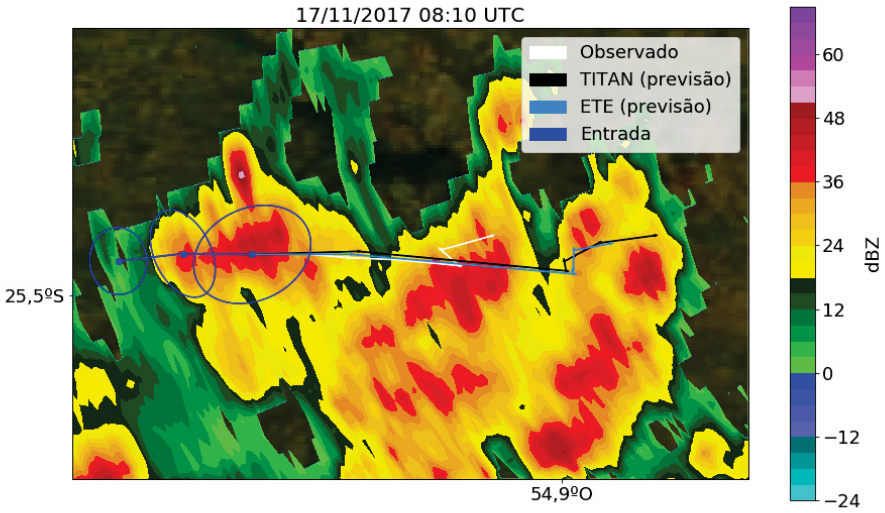
FONTE: A autora (2018)

FIGURA 31 – PREDIÇÃO DO MÉTODO RANDOM FOREST (17/11/2017 8:10 UTC)



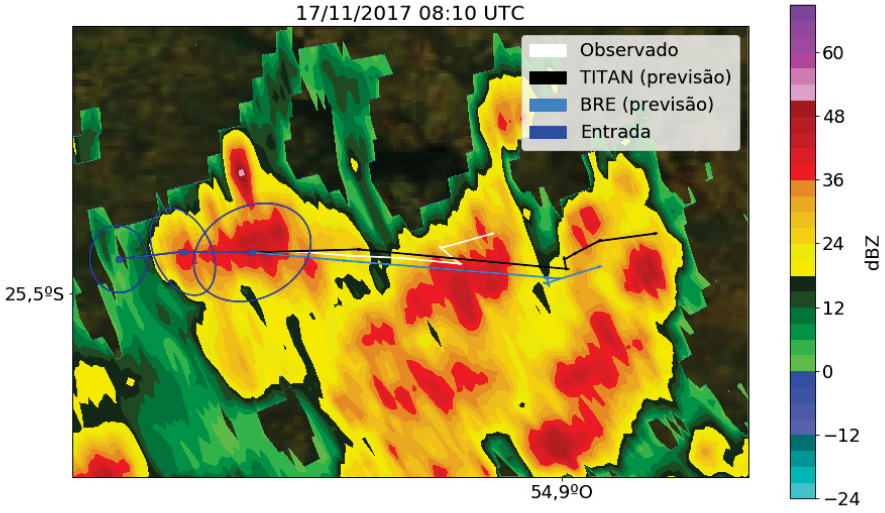
FONTE: A autora (2018)

FIGURA 32 – PREDIÇÃO DO MÉTODO EXTRA TREES (17/11/2017 8:10 UTC)



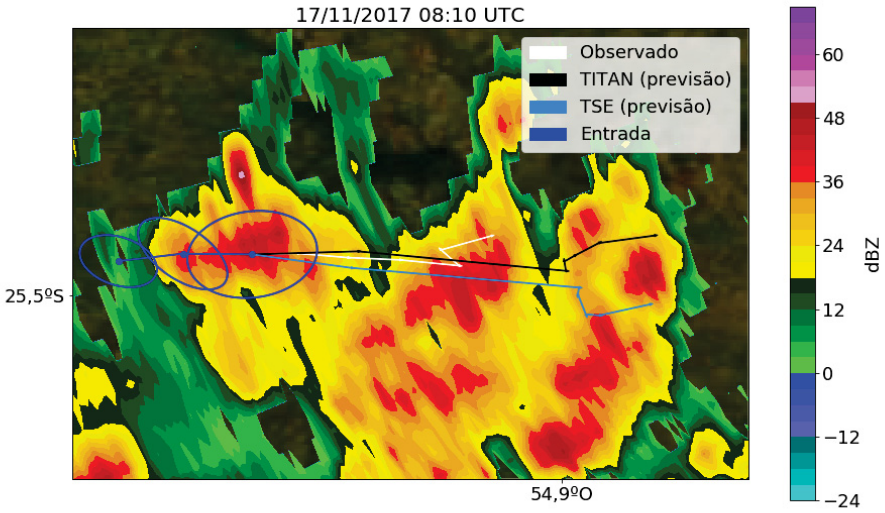
FONTE: A autora (2018)

FIGURA 33 – PREDIÇÃO DO MÉTODO BAYESIAN RIDGE (17/11/2017 8:10 UTC)



FONTE: A autora (2018)

FIGURA 34 – PREDIÇÃO DO MÉTODO THEIL SEN (17/11/2017 8:10 UTC)



FONTE: A autora (2018)

mento do centróide ao longo do tempo. Não é possível dizer que isso se repete para qualquer caso, mas essa análise condiz com os resultados apresentados, em que os métodos de aprendizado agrupado são capazes de descrever melhor os dados, quando comparados aos lineares robustos.

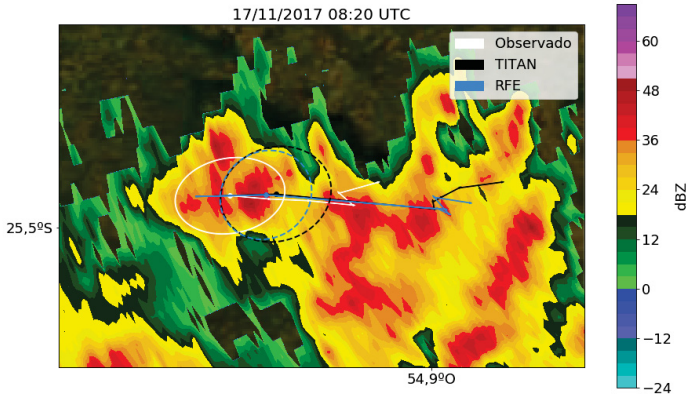
Todas as métricas apresentadas apontam para o modelo desenvolvido pelo algoritmo Random Forest (RFE) como a melhor técnica de AM ao problema de pesquisa. Portanto, o acompanhamento das previsões das tempestades ao longo de uma rota atém-se a esse algoritmo, conforme apresenta a FIGURA 35 para os primeiros 30 minutos previstos e a FIGURA 36 para os 30 restantes.

O evento selecionado também é interessante para visualizar os tamanhos das elipses previstas, em que mostra-se evidente a degeneração da área prevista pelo software para períodos de tempos maiores, enquanto o algoritmo de AM apresenta uma menor divergência.

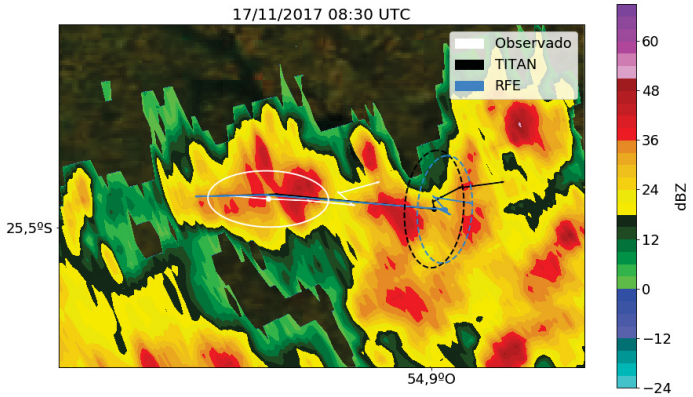
Foram apresentadas neste capítulo as avaliações dos desempenhos dos algoritmos, sendo a generalização, os métodos ensemble, os modelos lineres e uma análise do métodos de AM em comparação com o TITAN.

Com isso, propõe-se o algoritmo *Random Forest* como alternativa à predição obtida através do software TITAN e é reconhecida a contribuição das abordagens propostas para a estimativa do tamanho futuro da tempestade.

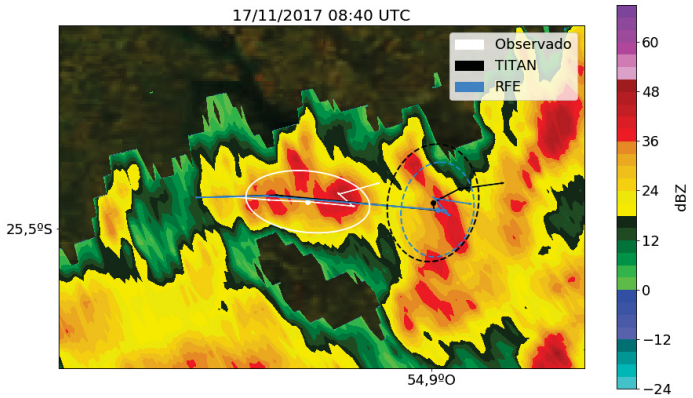
FIGURA 35 – PREDIÇÃO DO *RANDOM FOREST* E TITAN DE 10 A 30 MINUTOS (17/11/2017)



(a) T + 10 minutos



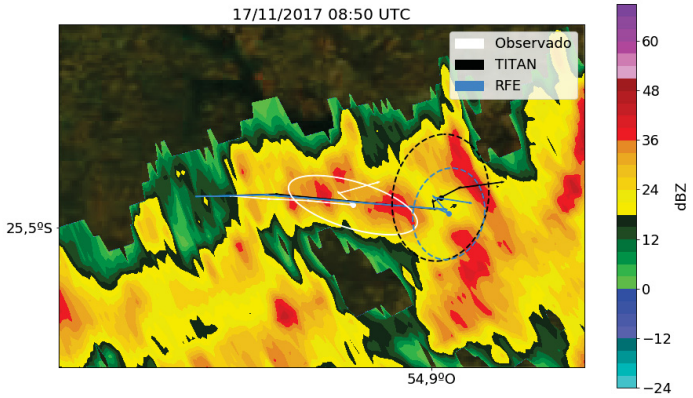
(b) T + 20 minutos



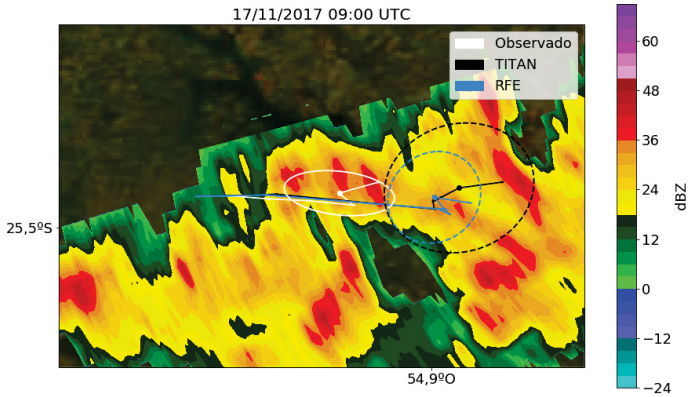
(c) T + 30 minutos

FONTE: A autora (2018)

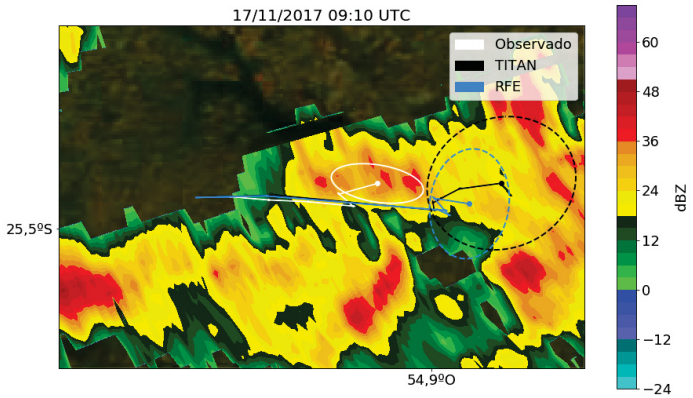
FIGURA 36 – PREDIÇÃO DO *RANDOM FOREST* E TITAN DE 40 A 60 MINUTOS (17/11/2017)



(a) T + 40 minutos



(b) T + 50 minutos



(c) T + 60 minutos

FONTE: A autora (2018)

6 CONCLUSÕES

As tempestades severas podem apresentar diversos riscos à economia e à vida humana. O acompanhamento desses eventos permite tomar as medidas possíveis e, por meio da previsão imediata do deslocamento, mitigar danos.

Os dados utilizados no estudo são de radares meteorológicos processados pelo TITAN, aliados às descargas elétricas atmosféricas, visto que são capazes que captar grande parte das características dos eventos atuantes. A seleção dos dados de forma recursiva permite a definição de um conjunto de dados que apresenta bons resultados para todos os modelos testados, conforme apresentado no QUADRO 3.

Os algoritmos propostos pertencem a dois grupos: o *ensemble*, composto pelos métodos *Bagging*, *Random Forest* e *Extra Trees* e o Modelo Linear, composto pelo *Theil Sen* e o *Bayesian Ridge*.

A avaliação de todos os métodos em um conjunto de dados totalmente externo aos utilizados nos conjuntos de aprendizagem (treinamento e teste) ilustram um grande potencial de generalização.

Segundo as mensurações dos erros, a eficácia dos algoritmos mantém uma ordem, indiferente da métrica avaliada, ou do atributo que está sendo previsto, destacando-se o RFE, seguido pelo ETE, TBE, BRE e, por último, o TSE. Assim, pode-se afirmar que as técnicas de aprendizado agrupado expressam melhor ajuste ao problema do que as de modelo linear, isso pode ser devido aos *ensembles* se tratarem de métodos não lineares e se ajustarem mesmo a dados ruidosos. Em comparação ao TITAN, apenas o *Theil Sen* exibe menor pertinência para a predição do deslocamento.

Os resultados da pesquisa são satisfatórios, uma vez que apontam erros de predição semelhantes e, em sua maioria, menores do que os obtidos pelo TITAN, técnica amplamente utilizada na área. A principal contribuição dos métodos de AM, quanto a estimativa desse atributo, é encontrada avaliando-se o erro médio absoluto, juntamente com a raiz do erro médio quadrático. O TITAN apresenta uma variação mais elevada entre as métricas, indicando uma discrepância maior entre as previsões.

Esse resultado se torna ainda mais evidente avaliando-se a expansão dos modelos para a previsão dos eixos da elipse de tempestade, bem como a área. Para esses atributos o benefício apresentado pelas técnicas de AM é realçado, uma vez que o TITAN se torna insuficiente para previsões maiores que 40 minutos.

Além disso, nessa configuração do problema o ganho no desempenho dos algoritmos propostos é evidenciado. A diferença nas métricas avaliadas se torna mais

destoante a cada período de tempo. Quando toma-se o EMA e REMQ, é possível observar que mesmo atribuindo maior peso à discrepância das previsões em torno do erro médio avaliado, ou seja, valiando-se a métrica REMQ para os modelos de AM, ainda se destacam em relação ao simples EMA do software TITAN, quanto mais a raiz do erro médio quadrático em si.

Apesar do objetivo principal da pesquisa ter por foco o deslocamento das tempestades, a maior contribuição dos algoritmos propostos é obtida na previsão do tamanho dos eixos da tempestade e, consequentemente, da sua área. O desempenho quanto ao deslocamento não deixa a desejar, mantendo-se melhor do que o software para a maioria dos métodos, mas em uma escala menor.

O sistema de previsão do TITAN existe a mais de três décadas e, durante este período de tempo, tem sido aprimorado. O SIMEPAR, como também outros orgãos no Brasil e em outros países, adotam o TITAN para realizar processo de previsão de tempestades severas. Por ser um método tão utilizado, optou-se por adotá-lo como referência para comparar resultados nesta área.

Dessa forma, os modelos mostraram-se não apenas suficientes mas, também, como uma alternativa promissora ao software TITAN. Portanto, propõe-se a adoção do *Random Forest*, método que exibe o melhor ajuste ao problema, dentre todos os estudados neste trabalho.

6.1 SUGESTÕES PARA TRABALHOS FUTUROS

A área de AM, apresenta uma gama de diferentes algoritmos que podem ser considerados para um mesmo problema, a partir de testes preliminares e revisão teórica foram selecionados dois grupos, totalizando cinco métodos. Em geral, esse tipo de abordagem tem por objetivo obter conhecimento por meio dos dados apresentados. Isso pressupõe um estudo exaustivo da configuração do algoritmo e dos atributos disponíveis.

Nesse sentido, outras técnicas ou formulação de Aprendizado de Máquina, como hibridização, podem ser estudadas. Com o intuito de verificar a previsão para mais posições ou uma acurácia maior à curto prazo.

O estudo desenvolvido, visou o deslocamento de tempestades, uma expansão interessante é verificar os algoritmos para a previsão de outros atributos, relativos a intensidade da tempestade, por exemplo. Visto que "para alguns clientes, uma imagem precisa do caminho anterior, extensão e gravidade das tempestades é tão importante ou até mais importante do que uma previsão para o futuro local da tempestade"(BALLY,

2001).

Os resultados encontrados para a predição do tamanho da célula de tempestade, são consequência de modelos construídos para fins relativos à variação da sua posição no tempo. Com isso, um estudo que objetiva a predição desse atributo pode encontrar avanços ainda mais satisfatórios.

Um ponto relevante na aplicação de AM, é a quantidade de dados de entrada e a qualidade desses dados. Ou seja, históricos mais completos para a células de tempestades ao longo do tempo, pode implicar em grandes melhorias na previsão, uma vez que o desempenho dessas técnicas está relacionado com a qualidade dos recursos aprendidos.

Portanto, estudos em geral para os processos de identificação e acompanhamento de tempestades podem contribuir para o avanço da previsão proposta. Mas, nesse momento, se propõe o estudo de abordagens de AM para essas outras duas etapas, como técnicas de clusterização, reconhecimento de imagens, entre outras.

Por fim, as técnicas de Aprendizado de Máquina podem ser consideradas para o problema proposto, uma vez que já apresentam resultados interessantes. Para um horizonte futuro, propõe-se o desenvolvimento de um sistema completo de acompanhamento a partir desse tipo de abordagem.

REFERÊNCIAS

ANOCHI, J. A.; CAMPOS VELHO, H. F. de. Previsão climática de precipitação para a região Sul por rede neural autoconfigurada. **Ciência e Natura**, v. 38, 2016.

ANOCHI, J. A.; SILVA, J. D. SD. Uso de teoria de conjuntos aproximativos e redes neurais artificiais no estudo de padrões climáticos sazonais. **Learning and Nonlinear Models**, v. 7, p. 83–91, 2009.

AUDHKHASI, K. et al. Creating ensemble of diverse maximum entropy models. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH e SIGNAL PROCESSING (ICASSP), 2012, p. 4845–4848.

BALLY, J. Generating severe weather warnings from TITAN and SCIT thunderstorm tracks. In: 30 TH CONFERENCE ON RADAR METEOROLOGY, 2001, Munich, Germany, p. 489–491.

BENETI, C. A. A. **Caracterização hidrodinâmica e elétrica de sistemas convectivos de mesoescala**. 2012. Tese (Doutorado) – Universidade de São Paulo.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. 1st ed. 2006. Corr. 2nd printing. [S.l.]: Springer, 2006. (Information science and statistics). ISBN 9780387310732.

BONATO, J. V. R. **Clusterização de dados meteorológicos para comparação de técnicas de nowcasting**. 2014. Diss. (Mestrado) – Universidade de Federal do Parana.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, n. 2, p. 123–140, 1996.

_____. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. et al. **Classification and regression trees**. [S.l.]: CRC, 1984. (The Wadsworth statistics / probability series). ISBN 0-412-04841-8.

BRILHADORI, M.; LAURETTO, M. S. Estudo comparativo entre algoritmos de árvores de classificação e máquinas de vetores suporte, baseados em ensembles de classificadores. **IX Simpósio Brasileiro de Sistemas de Informação**, 2013.

BROWN, G.; WYATT, J. L.; TIŇO, P. Managing diversity in regression ensembles. **Journal of machine learning research**, v. 6, Sep, p. 1621–1650, 2005.

CALHEIROS, A. J. P. **Sistema de previsão imediata da precipitação: o Hydrotrack**. 2008. Diss. (Mestrado) – Instituto Nacional de Pesquisas Espaciais.

COELHO, G. P. et al. Geração, seleção e combinação de componentes para ensembles de redes neurais aplicadas a problemas de classificação. [sn], 2006.

COENEN, F.; PREECE, A.; MACINTOSH, A. **Research and Development in Intelligent Systems XX: Proceedings of AI2003, the Twenty-third SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence**. [S.l.]: Springer Science & Business Media, 2011.

DAMIAN, E. A. **Duas metodologias aplicadas à classificação de precipitação convectiva e estratiforme com radar meteorológico**. 2012. Diss. (Mestrado) – Universidade de Federal do Parana.

DANG, X. et al. Theil-Sen Estimators in a Multiple Linear Regression Model. **Olemiss. edu**, Citeseer, 2008.

DESLANDES, R.; RICHTER, H.; BANNISTER, T. The end-to-end severe thunderstorm forecasting system in Australia: overview and training issues. **Australian Meteorological Magazine**, Bureau of Meteorology, v. 57, n. 4, 2008.

DIETTERICH, T. G. Ensemble methods in machine learning. In: INTERNATIONAL WORKSHOP ON MULTIPLE CLASSIFIER SYSTEMS, 2000, p. 1–15.

DIXON, M.; WIENER, G. TITAN: Thunderstorm identification, tracking, analysis, and nowcasting—A radar-based methodology. **Journal of atmospheric and oceanic technology**, v. 10, n. 6, p. 785–797, 1993.

FABRY, F. **Radar meteorology: principles and practice**. [S.l.]: Cambridge University Press, 2015.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine learning**, Springer, v. 63, n. 1, p. 3–42, 2006.

GOMES, Ana Maria; HELD, Gerhard. Identificação, rastreamento e previsão de tempestades severas. Parte I: Evento de granizo. In: XVI CONGRESSO BRASILEIRO DE METEOROLOGIA, SBMET, 2006, Florianópolis, SC.

GUILHON, L. G. F.; ROCHA, V. F.; MOREIRA, J. C. Comparação de métodos de previsão de vazões naturais afluentes a aproveitamentos hidroelétricos. **Revista Brasileira de Recursos Hıdricos**, v. 12, n. 3, p. 13–20, 2007.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. **Machine learning**, Springer, v. 46, n. 1-3, p. 389–422, 2002.

HAIR, J. F et al. **Análise multivariada de dados**. [S.l.]: Bookman Editora, 2009.

HAN, L. et al. 3D convective storm identification, tracking, and forecasting—An enhanced TITAN algorithm. **Journal of Atmospheric and Oceanic Technology**, v. 26, n. 4, p. 719–732, 2009.

HARRINGTON, P. **Machine learning in action**. [S.l.]: Manning Greenwich, CT, 2012. v. 5.

HAYKIN, S. **Redes neurais: princıpios e prática**. [S.l.]: Bookman Editora, 2007.

HEIDLER, F et al. Parameters of lightning current given in IEC 62305-background, experience and outlook. In: 29TH INTERNATIONAL CONFERENCE ON LIGHTNING PROTECTION, 2008. v. 23, p. 26.

HELMUS, J. J; COLLIS, S. M. The Python ARM Radar Toolkit (Py-ART), a Library for Working with Weather Radar Data in the Python Programming Language. **Journal of Open Research Software**, v. 4, n. 1, e25, 2016. ISSN 2049-9647. DOI: 10.5334/jors.119.

HERBRICH, R. **Learning Kernel Classifiers: Theory and Algorithms**. [S.l.]: The MIT Press, 2001. (Adaptive Computation and Machine Learning). ISBN 026208306X.

HERING, A et al. Nowcasting thunderstorms in the Alpine region using a radar based adaptive thresholding scheme. In: 6. PROCEEDINGS of ERAD. [S.l.: s.n.], 2004. v. 1.

HUBER, P. J.; RONCHETTI, E. M. **Robust statistics**. 2°. [S.l.]: Wiley, 2009. (Wiley Series in Probability and Statistics). ISBN 9780470129906,0470129905.

ILLINOIS. University of Illinois WW2010 Project., 2010. Disponível em: <http://ww2010.atmos.uiuc.edu/(Gh)/guides/rs/rad/basics/sgnl.rxml>. Acesso em: 17 jan. 2018.

INMET. **Instituto Nacional de Meteorologia**. [S.l.: s.n.]. Disponível em: <http://www.inmet.gov.br/portal/>. Acesso em: 17 ago. 2017.

INPE. **Instituto Nacional de Pesquisas Espaciais**. [S.l.: s.n.], 2015. Disponível em: <http://www.inpe.br>. Acesso em: 15 maio 2015.

_____. **Vítimas de Raios Infográfico**. Instituto Nacional de Pesquisas Espaciais: [s.n.]. Disponível em: <http://www.inpe.br/webelat/homepage/menu/noticias/vitimas.de.raios.-.infografico.php>. Acesso em: 17 ago. 2017.

JOHNSON, J. T et al. The storm cell identification and tracking algorithm: An enhanced WSR-88D algorithm. **Weather and forecasting**, v. 13, n. 2, p. 263–276, 1998.

KLEINA, M. **Identificação, monitoramento e previsão de tempestades elétricas**. 2015. Tese (Doutorado) – Universidade de Federal do Paraná.

KLEINA, M.; MATIOLI, L. C.; ALVIM LEITE, E. Identificação, monitoramento e previsão de tempestades elétricas utilizando métodos numéricos. **Boletim de Ciências Geodésicas**, Universidade Federal do Paraná, v. 22, n. 4, 2016.

KORONACKI, J.; RAS, Z. W; WIERZCHON, S. T. **Advances in Machine Learning II: Dedicated to the Memory of Professor Ryszard S. Michalski**. [S.l.]: Springer, 2009. v. 263.

KROGH, A.; VEDELSBY, J. Neural Network Ensembles, Cross Validation, and Active Learning. **Advances in Neural Information Processing Systems**, MIT Press, v. 7, p. 231, 1995.

LIBRALÃO, G. L. et al. Técnicas de Aprendizado de Máquina para análise de imagens oftalmológicas. **São Paulo. Universidade de São Paulo**, 2003.

LOHMANN, M. **Regressão logıstica e redes neurais aplicadas à previsão probabilıstica de alagamentos no Municıpio de Curitiba, PR**. 2011. Tese (Doutorado) – Universidade de Federal do Paraná.

LORENZETT, C. D. C.; TELÖCKEN, A. V. Estudo Comparativo entre os algoritmos de Mineração de Dados Random Forest e J48 na tomada de Decisão. In: SIMPÓSIO DE PESQUISA E DESENVOLVIMENTO EM COMPUTAÇÃO (SPDC), 1., 2016. v. 2.

LOUPPE, G. et al. Understanding variable importances in forests of randomized trees. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2013, p. 431–439.

MACKAY, D. J. C. Bayesian interpolation. **Neural computation**, MIT Press, v. 4, n. 3, p. 415–447, 1992.

MCKINNEY, W. pandas: a foundational Python library for data analysis and statistics. **Python for High Performance and Scientific Computing**, p. 1–9, 2011.

MOHAN, A.; CHEN, Z.; WEINBERGER, K. Web-search ranking with initialized gradient boosted regression trees. In: PROCEEDINGS OF THE LEARNING TO RANK CHALLENGE, p. 77–89.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. [S.l.]: MIT press, 2012.

NCAR, National Center for Atmospheric Research. **TITAN Thunderstorm Identification Tracking Analysis and Nowcasting**. [S.l.: s.n.], 2016. Disponível em: <http://www.ral.ucar.edu/projects/titan/>. Acesso em: 17 ago. 2017.

NEAL, R. M. **Bayesian learning for neural networks**. [S.l.]: Springer Science & Business Media, 2012. v. 118.

OLIVEIRA, C. de. **Identificação e correção da banda brilhante em dados de radar meteorológico**. 2014. Diss. (Mestrado) – Universidade de Federal do Parana.

OSHIRO, T. M. **Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica**. 2013. Tese (Doutorado) – Universidade de São Paulo.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PESSOA, A. S. A. **Mineração de dados meteorológicos pela teoria dos conjuntos aproximativos na previsão de clima por redes neurais artificiais**. 2004. Diss. (Mestrado) – Instituto Nacional de Pesquisas Espaciais.

PESSOA, A. S. A. **Predição de eventos severos em saídas de modelos meteorológicos utilizando a teoria dos conjuntos aproximativos e metaheurRísticas para redução de atributos**. 2014. Tese (Doutorado) – Instituto Nacional de Pesquisas Espaciais.

PESSOA, A. S. A. et al. Mineração de dados meteorológicos para previsão de eventos severos. **Revista Brasileira de Meteorologia**, SciELO Brasil, v. 27, n. 1, 2012.

QUEIROZ, A. P. de. **Monitoramento e previsão imediata de tempestades severas usando dados de radar**. 2009. Diss. (Mestrado) – Instituto Nacional de Pesquisas Espaciais.

RÄTSCH, G. A brief introduction into machine learning. **Friedrich Miescher Laboratory of the Max Planck Society**, 2004.

RAWLINGS, J. O; PANTULA, S. G; DICKEY, D. A. **Applied regression analysis: a research tool**. [S.l.]: Springer Science & Business Media, 2001.

RINDAT. **Rede Integrada Nacional de Detecção de Descargas Atmosférica**. [S.l.: s.n.]. Disponível em: <http://simepar.br/rindat/>. Acesso em: 11 jun. 2018.

RINEHART, R. E. **Radar for meteorologists**. 4. ed. [S.l.]: Rinehart Publishing, 2004.

SAFIER, F. **Pré-Calculo: Coleção Schaum**. [S.l.]: Bookman Editora, 2009.

SANTOS, T. N. dos. Redes neurais artificiais e relação ZR aplicadas à estimativa de chuva, 2014.

SEBER, G. A. F.; WILD, C. J. **Nonlinear Regression**. [S.l.]: Wiley, 2003. (Wiley series in probability and mathematical statistics. Probability and mathematical statistics). ISBN 9780471471356.

SEN, P. K. Estimates of the regression coefficient based on Kendall's tau. **Journal of the American statistical association**, Taylor & Francis Group, v. 63, n. 324, p. 1379–1389, 1968.

SHIGA, A. A. **Avaliação de custos decorrentes de descargas atmosféricas em sistemas de distribuição de energia**. 2007. Tese (Doutorado) – Universidade de São Paulo.

SILVA, A. L. C. Da. **Introdução à análise de dados**. [S.l.: s.n.], 2009.

SILVA, T. G. G. J. **Identificação de evento de tempo severo utilizando técnicas de Aprendizagem de Máquina em dados de radar polarimétrico**. 2017. Diss. (Mestrado) – Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná.

SOS-CHUVA, O. Previsão Imediata de Tempestades Intensas e Entendimento dos Processos Fısicos no Interior das Nuvens. INPE/CPTEC, 2015.

TEIXEIRA, M. A. de B. **Análise da trajetória e da circulação de sistemas precipitantes**. 2010. Diss. (Mestrado) – Instituto Nacional de Pesquisas Espaciais.

THEIL, H. A rank-invariant method of linear and polynomial regression analysis I. In: PROCEEDINGS OF KONINALIJKE NEDERLANDSE AKADEMIE VAN WEINENSCHATPEN A, 1950. v. 53, p. 386–392.

VILA, D. A. et al. Forecast and Tracking the Evolution of Cloud Clusters (ForTraCC) using satellite infrared imagery: Methodology and validation. **Weather and Forecasting**, v. 23, n. 2, p. 233–245, 2008.

WICHARD, J. D; OGORZALEK, M. Time series prediction with ensemble models. In: IEEE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, PROCEEDINGS, 2004. v. 2, p. 1625–1630.

WILK, K. E; GRAY, K. C. Processing and analysis techniques used with the NSSL weather radar system. In: 14TH CONFERENCE ON RADAR METEOROLOGY, 1970, Tucson, AZ, p. 369–374.

WILKS, D. S. **Statistical methods in the atmospheric sciences**. [S.l.]: Academic press, 2011. v. 100.

WILSON, J. W et al. Nowcasting thunderstorms: A status report. **Bulletin of the American Meteorological Society**, v. 79, n. 10, p. 2079–2099, 1998.